Wer profitiert von zentralen Abiturprüfungen?

Längerfristige Effekte der Implementation zentraler Abiturprüfungen im Bundesland Freie Hansestadt Bremen auf Handlungen und Emotionen von Lehrpersonen, Schülerinnen und Schülern

Abhandlung (kumulative Dissertation)
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der
Universität Zürich

vorgelegt von
Elisabeth Maué

Angenommen im Herbstsemester 2017
auf Antrag der Promotionskommission

Prof. Dr. Katharina Maag Merki (hauptverantwortliche Betreuerin)
Prof. Dr. Isabell van Ackeren

Zürich, 2018

Danksagung

Mein Dank gilt allen, die mich bei der Erstellung dieser Arbeit unterstützt haben.

Allen voran danke ich Frau Prof. Dr. Katharina Maag Merki ganz herzlich für viele inhaltliche Anregungen, die Möglichkeiten, Neues auszuprobieren und die Förderung meiner wissenschaftlichen Neugier während meiner Jahre in Zürich und darüber hinaus.

Frau Prof. Dr. Isabell van Ackeren danke ich ganz herzlich für ihre motivierenden Worte zwischendurch und die Übernahme der Zweitbetreuung.

Merci an meine (ehemaligen) Kolleginnen und Kollegen für den fachlichen Austausch, kontroverse Diskussionen und Gespräche auch abseits wissenschaftlicher Themen.

Darüber hinaus gebührt allen Beteiligten an der Zentralabitur-Studie, insbesondere den Lehrpersonen, Schülerinnen und Schülern, mein Dank. Ohne die Mitwirkung über viele Jahre hinweg wären die Analysen dieser Arbeit nicht möglich gewesen.

Meiner Familie sowie meinen Freundinnen und Freunden möchte ich für ihre Begleitung und Unterstützung von Herzen danken.

Zusammenfassung

Die Arbeit zeichnet für Bremen längerfristige Effekte (2007-2011) der Implementation zentraler Abiturprüfungen auf Lehrpersonen, Schülerinnen und Schüler nach. Das eigene Modell der Wirkungen der Implementation des Zentralabiturs basiert auf der Mehrebenenstruktur des Bildungssystems, Rekontextualisierungsprozessen verschiedener Akteure, der Steuerung des Bildungssystems, der Handlungskoordination unterschiedlicher Akteure im Mehrebenensystem, der Implementation von Reformen und Innovationen, Schulentwicklung und Schuleffektivität. Die Befunde verweisen darauf, dass die erhöhte Standardisierung der Abiturprüfungen und deren Benotung die Vergleichbarkeit der Abiturnoten in Mathematik-Leistungskursen nicht steigert. Bei den Halbjahresnoten lassen sich hingegen einzelne Standardisierungseffekte ausmachen. Zudem reduzieren sich Unsicherheit, Leistungsdruck und Arbeitsunzufriedenheit der Lehrpersonen und deren Gefühl der Entlastung nimmt zu. Das komplexe Zusammenspiel von Emotionen der Schülerinnen und Schüler mit individuellen Merkmalen sowie unterrichtlichen und schulischen Faktoren bleibt über die Zeit hinweg relativ stabil. Allerdings verbessert sich die Unterrichtsqualität nach dem Urteil der Schülerinnen und Schüler. Einzelne spezifische Effekte deuten darauf hin, dass gewisse Akteure von der Implementation zentraler Abiturprüfungen profitieren, andere nicht. Insgesamt lässt sich jedoch kein genereller Effekt des Zentralabiturs auf das Lehren und Lernen erkennen.

Abstract

This thesis deals with the long-term effects (2007-2011) on teachers and students affected by the implementation of state-wide exit examinations in Bremen. It includes a new model showing potential changes caused by the implementation of state-wide exit examinations. This model is based on the multi-level structure of the educational system, the recontextualization processes of different parties, the steering of the educational system, the coordination of action of different parties in the multi-level system, the implementation of reforms and innovations, the school improvement as well as the school effectiveness. The findings prove that the raised standardization of state-wide exit examinations and their grading does not increase the comparability of exit examination grades in mathematics courses at advanced levels. However, in regards to semester grades, individual standardization effects can be observed. While uncertainty, performance pressure and job dissatisfaction of teachers is reduced, their stress relief increases. The interaction between the students' emotions and the individual characteristics, didactic and school factors stays relatively stable. However, according to students' opinion, the quality of teaching has improved. Specific effects indicate that certain parties benefit from the implementation of state-wide exit examinations while others do not. Overall, one cannot directly observe the general impact of state-wide exit examinations on teaching and learning.

Gliederung

Abbildungsverzeichnis

Veränderungen im Bildungssystem brauchen Zeit, viel Zeit – und während dieser Zeit Verlässlichkeit in der Orientierung und in den Rahmenbedingungen.

(Meyer-Hesemann, 2010, S. 87)

Getting a new idea adopted, even when it has obvious advantages, is difficult. Many innovations require a lengthy period of many years from the time when they become available to the time when they are widely adopted.

(Rogers, 2003, S. 1)

1. Einleitung

Laut einer aktuellen Studie befürworten in Deutschland 91% der Befragten einheitliche Abschlussprüfungen beim Abitur und zu einem ähnlich hohen Anteil einheitliche Abschlussprüfungen am Ende der Sekundarstufe I (Wößmann, Lergetporer, Grewenig, Kugler, & Werner, 2017, S. 28). Damit sind zentrale Abschluss- und Abiturprüfungen in der Breite akzeptiert und zwar vermutlich unabhängig davon, ob sie bereits auf eine lange Tradition zurückblicken oder wie in einigen Bundesländern erst seit etwa 10 bis 15 Jahren durchgeführt werden. Die Umstellung der Prüfungsorganisation von dezentralen, von Lehrpersonen erstellten zu zentralen Abiturprüfungen reiht sich in mehrere grundlegende „Transformationsprozesse" des Gymnasiums und des Abiturs ein (Neumann, 2014, S. 245; für einen Überblick über Reformen seit 1949 in den einzelnen Bundesländern siehe Helbig & Nikolai, 2015).

Mit Rückgriff auf theoretische Überlegungen zu „washback" bzw. „backwash" ist davon auszugehen, dass Prüfungen und Tests das Lehren und Lernen beeinflussen (Amengual Pizarro, 2010; Bishop, 1995; Cheng, Watanabe, & Curtis, 2004; Haertel, 2013; Prodromou, 1995; Scott, 2011). Bisherige Forschungsbefunde zu Effekten der Einführung zentraler Abiturprüfungen differieren nach Ländern, Bundesländern, Jahrgangsstufen, Schulformen, Fächern, Kursniveaus und Zeiträumen. Insgesamt ist somit nicht von einem generellen Effekt des Zentralabiturs auf das Lehren und Lernen auszugehen. Da die bisherigen Studien jedoch meist lediglich einen kurzen Zeithorizont umfassen, können keine Aussagen zu längerfristigen Effekten und Entwicklungen getroffen werden. Diesem Desiderat begegnet die vorliegende Arbeit, indem sie Auswirkungen auf die im besonderen von der Reform des Abiturs betroffenen Akteure – die unterrichtenden Lehrpersonen sowie die Schülerinnen und Schüler – in längerfristiger Perspektive über einen Zeitraum von fünf Jahren in den Blick nimmt. Die übergeordneten Fragestellungen umfassen zum einen von Seiten der Bildungspolitik intendierte und nicht-intendierte Auswirkungen der Reform und zum anderen „äussere" und „innere" Effekte. Es wird erstens gefragt, ob die Implementation zentraler Abiturprüfungen zu einer Steigerung der Vergleichbarkeit der Abiturnoten und der Halbjahresnoten im Fach Mathematik führt (intendierter, äusserer Effekt). Zweitens verweisen die Fragestellungen nach dem emotionalen Erleben des Zentralabiturs von Lehrpersonen sowie von Schülerinnen und Schülern auf innere, intendierte Effekte (z. B. Entlastung der Lehrpersonen) und innere, nicht-intendierte Effekte (Erhöhung von Druck und Stress). Basierend auf der Annahme differenzieller Auswirkungen der

Einführung des Zentralabiturs auf unterschiedliche Akteure wird drittens der Frage nachgegangen, ob sich Akteure abzeichnen, die davon profitieren bzw. nicht profitieren.

Die Herleitung und Beantwortung der Fragestellungen basiert auf vielfältigen theoretischen Bezügen zu den Handlungen verschiedener Akteure im Mehrebenensystem Schule und ist in die bisherige Forschung zu den Wirkungen zentraler Abiturprüfungen eingebettet.

Originalität

Die vorliegende Arbeit steht in Ergänzung zu und Erweiterung von Fragestellungen, die im Rahmen des Projektes „Implementierung und Auswirkungen neuer Steuerungsstrukturen im Schulwesen am Beispiel zentraler Abiturprüfungen. Eine Analyse der Effekte unter Berücksichtigung multipler Indikatoren" (Maag Merki, 2012c) bearbeitet wurden. Sie richtet den Blick auf das Bundesland Freie Hansestadt Bremen[1]. Die Verlängerung des Zeithorizonts mit einer zusätzlichen Datenerhebung im Jahr 2011 ermöglicht die Analyse von Effekten über einen *Zeitraum* von fünf Jahren. Die schrittweise Implementation zentraler Abiturprüfungen begründet eine Besonderheit in der Datenlage für die Leistungskurse: Es liegen sowohl *Daten* zum Zeitpunkt der Einführung zentraler Abiturprüfungen vor (ab 2008) als auch solche vor der Reform, das heisst unter den Bedingungen dezentraler Abiturprüfungen (2007). So können nicht nur kurzfristige Effekte untersucht werden, die in direktem Zusammenhang mit dem Wechsel des Prüfsystems stehen, sondern es können auch diese Effekte in längerfristiger Perspektive (2007 bis 2011) hinsichtlich Stabilität oder weiterer Entwicklungen eingeordnet werden. Die Datenbasis ermöglicht als *Auswertungsstrategien* je nach Fragestellung hierarchisch lineare Modelle zur Berücksichtigung der Mehrebenenstruktur oder Strukturgleichungsmodelle.

Zudem berücksichtigt die Arbeit unterschiedliche *Akteure*: Lehrpersonen im Quer- und im Längsschnitt sowie Schülerinnen und Schüler in Leistungskursen (fachunabhängig) und fachspezifisch im Leistungskurs Mathematik. Inhaltlich richtet sich der Fokus auf das „Äussere", das in der *Vergleichbarkeit von Noten* sichtbar wird, wie auch auf das „Innere", das im *emotionalen Erleben* der Akteure verortet ist. Damit einhergehend werden zentralen Abschluss- und Abiturprüfungen zugeschriebene Wirkungen wie die bildungspolitisch intendierte Vergleichbarkeit von Noten aufgrund höherer Standardisierung und Entlastung der Lehrpersonen (Die Senatorin für Bildung und Wissenschaft, 2013a) sowie Kritikpunkte

[1] im Folgenden Bremen

von nicht-intendiertem erhöhten Stress und Druck für die beteiligten Akteure aufgegriffen (Amrein & Berliner, 2002a; Bishop, 1999; Pedulla et al., 2003; van Ackeren, Block, Klein, & Kühn, 2012).

Der spezifische Blick auf das Bundesland Bremen ermöglicht die Analyse längerfristiger Effekte des Wechsels von dezentralen zu zentralen Abiturprüfungen *innerhalb* eines Bundeslandes bei relativ konstanten Kontextbedingungen (Helbig & Nikolai, 2015; Klein & van Ackeren, 2011). Damit erweitert die vorliegende Arbeit bisherige Studien, die sich auf den Vergleich von Ländern oder Bundesländern mit und ohne zentrale Abschluss- bzw. Abiturprüfungen stützen (Baumert & Watermann, 2000; Jürges & Schneider, 2010; Jürges, Schneider, Senkbeil, & Carstensen, 2009; Wößmann, 2003).

> Sowohl der innerdeutsche als auch der internationalen [sic!] Vergleich zwischen zentral und dezentral prüfenden Systemen liefert keine überzeugenden Belege für einen eindeutigen generellen Zusammenhang zwischen Prüfungsform und erzielter Leistung (van Ackeren & Bellenberg, 2004, S. 147).

Die Analysen innerhalb eines Bundeslandes vertiefen das Verständnis der Auswirkungen der Implementation zentraler Abiturprüfungen und des Zusammenhangs von Prüfungsform und einem breiteren Begriff von „Leistungen" verschiedener Akteure.

Aufbau der Synopse

Der Einordnung der vorliegenden Arbeit in das Feld zentraler Abschluss- und Abiturprüfungen dient Kapitel 2 mit einem zunächst vorrangig international ausgerichteten Überblick sowie der anschliessenden Darstellung zentraler Abiturprüfungen in Deutschland (Kapitel 2.1) und speziell im Bundesland Bremen (Kapitel 2.2).

Anschliessend werden als theoretische Grundlage der Arbeit Rekontextualisierungsprozesse im Mehrebenensystem Schule (Kapitel 3.1), (Neue) Steuerung (Kapitel 3.2), Educational Governance (Kapitel 3.3), Reformen und Innovationen und deren Implementation (Kapitel 3.4) sowie Schulentwicklung und Schuleffektivität (Kapitel 3.5) beschrieben. Diese theoretischen Anknüpfungspunkte liefern den Rahmen für die Ableitung eines theoretischen Modells längerfristiger Effekte der Implementation zentraler Abiturprüfungen (Kapitel 3.6).

Die Aufbereitung empirischer Forschungsergebnisse zu den Wirkungen zentraler Abiturprüfungen auf Lehrpersonen (Kapitel 4.1), Schülerinnen und Schüler (Kapitel 4.2) sowie auf schulische Prozesse und den Unterricht (Kapitel 4.3) verweist auf das bestehende Forschungsdesiderat (Kapitel 4.4), dem die vorliegende Arbeit mit ihren übergeordneten Fragestellungen und Hypothesen (Kapitel 5) begegnet. Zusammenfassungen der Publikationen zur Vergleichbarkeit der Abiturnoten in Mathematik (Kapitel 6.1) und der Halbjahresnoten in Mathematik (Kapitel 6.2) sowie zum emotionalen Erleben des Zentralabiturs von Lehrpersonen (Kapitel 6.3) und Schülerinnen und Schülern (Kapitel 6.4) liefern die Grundlage zur Beantwortung der Forschungsfragen (Kapitel 7.1) sowie zur theoretischen Einordnung der Befunde (Kapitel 7.2). Anschliessend werden Limitationen aufgezeigt (Kapitel 7.3) sowie Implikationen für Forschung und Praxis abgeleitet (Kapitel 7.4). Ein Fazit rundet die Arbeit ab (Kapitel 8).

2. Zentrale Abschluss- und Abiturprüfungen

Weltweit absolvieren Schülerinnen und Schüler verschiedener Schulstufen zentrale Abschlussprüfungen. Trotz teilweise sehr unterschiedlicher Ausgestaltung (Klein, Kühn, van Ackeren, & Block, 2009; Klein & van Ackeren, 2011) umfassen die damit verbundenen *Ziele* vorrangig eine Erreichung, Sicherung und Steigerung der Leistungen sowie der Transparenz, Objektivität und Vergleichbarkeit von Anforderungen, Bewertungen, Noten und Abschlüssen und damit insgesamt der Qualität (Amrein & Berliner, 2002b; Bishop & Wößmann, 2004; Kühn, 2010, 2012; Maag Merki, 2008, 2016; Neumann, 2014; van Ackeren & Bellenberg, 2004; van Ackeren, Klemm, & Kühn, 2015; vbw – Vereinigung der Bayerischen Wirtschaft e. V., 2011). Des Weiteren wird angenommen, dass zentrale (Abschluss-)Prüfungen zur Sicherung von Standards und einer fundierten Grundbildung sowie zur Verbreitung von Innovationen beitragen (Kühn, 2010, S. 45). Zudem können sie das Monitoring der Leistungen von Schülerinnen und Schülern, Lehrpersonen und Schulen erleichtern (Bishop & Wößmann, 2004, S. 26).

Je nachdem, ob bzw. welche *Konsequenzen* mit den Prüfungsergebnissen einhergehen, wird zwischen low-stakes und high-stakes Prüfungen differenziert. Ein prominentes Beispiel für high-stakes Verfahren bilden die zentralen Tests, die in Amerika spätestens im Zuge des „No Child Left Behind-Aktes (NCLB)" flächendeckend eingesetzt werden. Unterdurchschnittliche Ergebnisse ziehen nicht nur für Schülerinnen und Schüler, sondern auch für Lehrpersonen, Schulleitungen und regionale Behörden Sanktionen bis hin zu Schulschliessungen nach sich (Amrein & Berliner, 2002b; Firestone & Schorr, 2004; Flaitz, 2011; Forte, 2010; Gogolin, Baumert, & Scheunpflug, 2011; Koretz, 2011; Madaus, Russell, & Higgins, 2009; Natriello, 2009; Nichols, Glass, & Berliner, 2006; Ravitch, 2010; für England: Gogolin, Baumert, & Scheunpflug, 2011, S. 2; Mansell, 2011). „Low-stakes testing, by contrast, has a reputation for being softer and less judgemental" (Allen, 2012, S. 641). Es bietet weniger Anreize für Lehrpersonen, Schülerinnen und Schüler, sich vorrangig auf ausgewählte, überprüfte Inhalte zu konzentrieren (Koretz, 2008b, S. 786). Bei low-stakes Verfahren wird zentralen Abschlussprüfungen eine Entlastung der Lehrpersonen zugeschrieben, da diese nicht für die Erstellung der Prüfungsaufgaben verantwortlich sind (Böhm-Kasper & Weishaupt, 2002; Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4; Klein et al., 2009, S. 620; Kühn, 2010; Maag Merki, 2008; van Ackeren & Bellenberg, 2004, S. 135). Lehrpersonen kennen ebenso wie ihre Schülerinnen und Schüler die Prüfungsaufgaben nicht, sodass sich deren Rolle (Begleitende,

Verbündete oder Coaches statt Richtende) sowie das Verhältnis der beiden Akteursgruppen zueinander anders ausgestalten kann als bei dezentralen, von den Lehrpersonen verantworteten Prüfungen (Maag Merki, 2008; van Ackeren et al., 2012; vbw – Vereinigung der Bayerischen Wirtschaft e. V., 2011).

Kritik

Zentrale Abschlussprüfungen sind nicht unumstritten. Besonders bei high-stakes Verfahren setzt die Kritik an mehreren Punkten an, die jedoch auch Prüfungen mit low-stakes Charakter betreffen.

Madaus et al. (2009, S. 141; ähnlich Koretz, 2008a; 2008b: „reallocation") unterscheiden drei Arten, wie sich high-stakes Tests auf das Lehren und Lernen auswirken können: Es wird erstens befürchtet, dass *innerhalb einer Disziplin* in den getesteten Fächern die Unterrichtsinhalte auf prüfungsrelevante Themen und Aufgabenformate eingegrenzt werden und dass damit eine Abkehr von Aufgaben, die Problemlösungen, Kreativität und Transfer von Gelerntem erfordern, einhergeht. Zudem besteht die Sorge, dass thematische Schwerpunktsetzungen basierend auf den Interessen der Lehrpersonen, Schülerinnen und Schüler oder auf aktuellen Ereignissen vernachlässigt werden. Unter diesen Bedingungen könnten Schülerinnen und Schüler nur einen eingeschränkten Ausschnitt ihrer Fähigkeiten und Kenntnisse zeigen (Bishop, 1999, S. 349; Firestone & Schorr, 2004, S. 2; Herman, 2005, S. 2; Koretz, 2008a; 2008b, S. 778ff.; Natriello, 2009, S. 1108f.; van Ackeren & Bellenberg, 2004, S. 134f.; van Ackeren et al., 2015, S. 177). Da meistens lediglich eine begrenzte Anzahl von Fächern zentral geprüft wird, bezieht sich die Kritik auch auf die Beeinflussung *unterschiedlicher Fächer* und die Befürchtung, dass die übrigen Fächer weniger Aufmerksamkeit, Zeit und Bedeutung erhalten (Koretz, 2008b, S. 781; Madaus et al., 2009; Natriello, 2009, S. 1108). Zum Dritten können sich zentrale Abschlussprüfungen auf *verschiedene Klassenstufen* auswirken, sofern zentrale Tests und Prüfungen Inhalte abfragen, die auf Wissen basieren, das über Jahre aufgebaut wurde. In diesem Fall wirken sie auch auf untere Klassenstufen, die nicht von zentralen Tests und Prüfungen betroffen sind, und verändern das dortige Curriculum (Madaus et al., 2009, S. 141).

Derartige und weitere Verschiebungen des Fokus in der *Organisation und Gestaltung von Schule und Unterricht* werden häufig unter dem Begriff „teaching to the test" subsumiert.

> This is a term with many meanings, but the implication is, generally speaking, that teachers are doing something special to help students do well on a test, often without helping them to better understand the underlying subject matter (Firestone & Schorr, 2004, S. 2).

Statt von *teaching to the test* zu sprechen, schlagen Hamilton, Stecher und Klein (2002, S. 87ff.) sowie Koretz (2008a, S. 251) vor, sieben Formen von „test preparation" zu unterscheiden: "Working more effectively", "teaching more" und "working harder" erhoffen sich Befürworter von high-stakes Tests, wohingegen „cheating" – beispielsweise mittels „providing answers or hints to students during the administration of the test, allowing students to change their answers after the test has been completed, changing the answers for them, providing test questions in advance" (Koretz, 2008a, S. 252) – indiskutabel ist und nicht zu Lernzuwachs führt. Die Umverteilung institutioneller Ressourcen („reallocation"), die Verbindung von Tests, Standards und Unterricht („alignment") sowie die Ausrichtung des Unterrichts auf begrenzte Ausschnitte des Tests („coaching") können sowohl in Lernzuwachs als auch in „score inflation" resultieren (Koretz, 2008a, S. 251ff.; 2008b, S. 781ff.). *Score inflation* oder *grade inflation* bezeichnet eine Steigerung der Testwerte und damit eine Verbesserung des Abschneidens einer Schule mit Hilfe von *teaching to the test*- bzw. *test preparation*-Strategien, jedoch ohne damit einhergehende Lernfortschritte der Schülerinnen und Schüler (Madaus et al., 2009, S. 3). Durch Einbezug der Lehrpersonen in die Korrektur zentraler Tests und Prüfungen könnten deren „taktische Verbesserungsmaßnahmen" (van Ackeren, 2005, S. 28) möglicherweise verringert werden.

An die Rolle und Bedeutung der *Lehrpersonen* anschliessend, besteht zudem die Befürchtung, dass durch starke Eingriffe in deren Handeln die Lehrpersonen deprofessionalisiert werden. Durch den Verlust der Attraktivität und Qualität des Lehrberufs würde das gesamte Bildungssystem geschwächt (Bellmann, 2016, S. 24; Bishop, 1999, S. 389f.; Black, Harrison, Hodgen, Marshall, & Serret, 2011; Bormann, 2012; Gehrmann, 2003, S. 31; Good, Wiley, & Sabers, 2010, S. 146; Madaus et al., 2009, S. 3; Natriello, 2009, S. 1109; van Ackeren et al., 2012, S. 15). Inwiefern in Deutschland zentrale Tests und Prüfungen Menschen davon abhalten, den Lehrberuf zu ergreifen, muss an dieser Stelle offen bleiben.

Da zentrale Abschlussprüfungen in Deutschland low-stakes Verfahren sind (Fend, 2011, S. 17f.)[2], dürfte dieser Kritikpunkt ein geringeres Gewicht haben als in high-stakes Kontexten. Allerdings kann erstens bei auf der Makroebene konzipierten low-stakes Verfahren durch entsprechende Strukturen und Handlungen auf der Mesoebene durchaus ein high-stakes Charakter entstehen. Zweitens muss zwischen objektiven und subjektiven stakes (deren Rezeption) differenziert werden (Bellmann, 2016, S. 25).

Schulen und Lehrpersonen haben die Aufgabe, Schülerinnen und Schüler mit unterschiedlichen Hintergründen unter verschiedenen Bedingungen auf dieselben Prüfungen vorzubereiten. Zentrale Prüfungen bieten demnach keinen Raum, „unterschiedliche Bedingungen der Leistungserbringung in den Schulen" (van Ackeren & Bellenberg, 2004, S. 135) zu berücksichtigen, was die Frage nach Chancengerechtigkeit aufwirft (Camilli & Monfils, 2004; Firestone & Schorr, 2004, S. 9f.; Koretz, 2011, S. 12f.; Madaus & Clarke, 2001, S. 20; Madaus et al., 2009; Monfils et al., 2004, S. 37). Empirische Befunde verweisen auf einen erhöhten Anteil von *Schülerinnen und Schülern*, welche eine Klasse wiederholen müssen oder die Schule ohne einen Abschluss verlassen. Darüber hinaus wirken sich negative Folgen der Tests nicht auf alle Schülerinnen und Schüler gleich aus, sondern haben insbesondere für jene aus Familien mit einem niedrigen sozialen Status und/oder einer Migrationsgeschichte einen stärkeren nachteiligen Effekt (Amrein & Berliner, 2002b, S. 10ff.; Madaus & Clarke, 2001; Natriello, 2009, S. 1106; Solórzano, 2008).

Nicht nur die Folgen zentraler Tests, sondern auch deren *Design* stehen unter kritischer Beobachtung. Vor dem Hintergrund des *No Child Left Behind*-Aktes in Amerika unterscheiden Good et al. (2010, S. 146) drei Ebenen von Lernzuwachs – individuell innerhalb einer Kohorte oder im Längsschnitt, auf Klassenebene als Mischung von Quer- und Längsschnitt und über die Zeit als Mischung von Kohorten- und Querschnittsvergleich (Koretz, 2011, S. 13: zusätzlich „static standards: comparison with a cut score fixed over time"). Sie kritisieren, dass die den Vergleichen zugrunde liegenden testtheoretischen Annahmen oftmals nicht eingelöst werden und kommen zu dem Urteil „that state-developed tests are not designed to measure growth" (Good et al., 2010, S. 147; auch Amrein & Berliner, 2002b; Hoover, 2014: „pseudoaccountability"; Koretz, 2011; Madaus et al., 2009; Wiliam, 2010). Hieran schliesst die vielfältige empirische Untersuchung von *score inflation* und *grade inflation* an, die mittlerweile ebenfalls in Deutschland stattfindet (z. B. Holcombe, Jennings, & Koretz, 2013; Koretz, 2008a; Koretz, 2008b, 2011; OECD, 2012; Sunderman, 2013; Wikström, 2005; für Deutschland: Grözinger & Baillet, 2015; Müller-Benedict & Grözinger, 2017; Wissenschaftsrat, 2012).

[2] Zu unterschiedlichen Sichtweisen siehe Hahn (2014, S. 41).

Nicht alle Kritikpunkte sind ausschliesslich negativ zu bewerten, sondern können in gewissem Mass auch positive Aspekte beinhalten. Insbesondere dem *teaching to the test* bzw. der *test preparation* wird die Möglichkeit eines „Korrektivs", das den Fokus von Lehrpersonen, Schülerinnen und Schülern auf die zentralen Aspekte des Curriculums lenkt, zugeschrieben. Richten Lehrpersonen ihren Unterricht an den getesteten Inhalten aus, können sie nicht nur die geforderten Inhalte in der gewünschten Breite und Tiefe abdecken und das Niveau gegebenenfalls steigern, sondern auch ihre Schülerinnen und Schüler auf den Ablauf standardisierter Tests vorbereiten (Firestone & Schorr, 2004, S. 2f.; Natriello, 2009, S. 1108; van Ackeren & Bellenberg, 2004, S. 135f.; van Ackeren et al., 2015, S. 177). In Deutschland bieten beispielsweise die mehr als zwei Jahre vor Durchführung zentraler Abiturprüfungen bekanntgegebenen Prüfungsschwerpunkte bzw. -themen nicht nur den Lehrpersonen, sondern auch den Schülerinnen und Schülern eine inhaltliche Orientierung.

Insgesamt ist von einer Mischung aus positiven und negativen Folgen sowie von „Grauzonen" zentraler Tests und Prüfungen, insbesondere mit high-stakes Charakter, auszugehen. Dabei dominiert eher eine pessimistische Sicht, wonach die negativen Auswirkungen überwiegen. „And over time, these negatives corrupt the accuracy of information about student achievement and school quality" (Madaus et al., 2009, S. 3; auch Amrein & Berliner, 2002, S. 11). Werden in Deutschland zentrale Abschluss- und Abiturprüfungen zu Monitoringzwecken eingesetzt, ist diese Kritik zu bedenken.

2.1 Zentrale Abiturprüfungen in Deutschland

Das Abitur bildet den Abschluss der Sekundarstufe II und bescheinigt den Schülerinnen und Schülern die allgemeine Hochschulreife und damit die Berechtigung zum Studium an Hochschulen und Universitäten. Mittlerweile werden in allen Bundesländern, mit Ausnahme von Rheinland-Pfalz, zentrale Abiturprüfungen mit unterschiedlicher Ausgestaltung durchgeführt. Die Einführung erfolgte deutschlandweit in drei Wellen. Nach dem Zweiten Weltkrieg übernahmen Baden-Württemberg, Bayern, Hessen, Rheinland-Pfalz und das Saarland zentrale Prüfungen von den Amerikanischen und Französischen Besatzungsmächten. Hessen und Rheinland-Pfalz schafften diese nach Gründung der Bundesrepublik Deutschland Anfang der 1950er Jahre wieder ab. Nach der deutschen Wiedervereinigung

behielten Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt und Thüringen die bereits in der Deutschen Demokratischen Republik bestehenden zentralen Prüfungen. Einzig Brandenburg schaffte sie zunächst ab. Im Rahmen umfangreicher Reformen nach PISA 2000 implementierten ab Mitte der 2000er Jahre Berlin, Brandenburg, Bremen, Hamburg, Hessen, Niedersachsen, Nordrhein-Westfalen und Schleswig-Holstein zentrale Abiturprüfungen. (Helbig & Nikolai, 2015; van Ackeren et al., 2015; vbw – Vereinigung der Bayerischen Wirtschaft e. V., 2011)

Organisation

Aufgrund des föderalen Systems gehört das Bildungssystem in Deutschland zu den Hoheitsrechten der einzelnen Bundesländer (Art. 30, 70 GG). In einigen Teilbereichen kann der Bund Vorgaben machen bzw. mit den Ländern kooperieren (konkurrierende Gesetzgebung, Art. 74 Abs. 1, 91b GG; ausführlicher van Ackeren et al., 2015, S. 95ff.). Die Bundesländer sind verantwortlich für die „Steuerung der eigentlichen *Unterrichts- und Erziehungsarbeit*" („innere Schulangelegenheiten"; van Ackeren et al., 2015, S. 99; Hervorhebung im Original). Bezüglich des Abiturs bedeutet dies, dass die Bundesländer die Prüfungsmodalitäten festlegen auf der Grundlage der „Vereinbarung über die Abiturprüfung der gymnasialen Oberstufe in der Sekundarstufe II" der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (2013). Zwischen den einzelnen Bundesländern bestehen nicht nur deutliche Differenzen bezüglich Anzahl und Art der Prüfungsfächer, sondern auch dahingehend, in wie vielen und in welchen Fächern auf welchem Anforderungsniveau zentrale Abiturprüfungen durchgeführt werden (Helbig & Nikolai, 2015; Kühn, 2012; Neumann, 2014; Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2013). Die mündlichen Abiturprüfungen erfolgen nach wie vor dezentral. Mittlerweile kooperieren einige Bundesländer beim Abitur („Kernabitur"; „Südabitur"; „Länderübergreifendes Abitur"; Kühn, 2012; Maag Merki, 2012b; Neumann, 2014; vbw – Vereinigung der Bayerischen Wirtschaft e. V., 2011). So entschieden sich Bayern, Brandenburg, Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Sachsen und Schleswig-Holstein für gemeinsame Prüfungsaufgaben in den Fächern Deutsch, Englisch und Mathematik.[3]

Als Basis für die *Prüfungsaufgaben* dienen die „Einheitlichen Prüfungsanforderungen in der Abiturprüfung" (EPA) der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik

[3] In Bremen werden die länderübergreifenden Aufgaben ab 2016 in Deutsch und Mathematik sowie ab 2017 in Englisch in die Abiturprüfungen integriert.
https://www.bildung.bremen.de/laenderuebergreifendes_abitur-116627 [27.07.2016]

Deutschland (2008). Sie enthalten „Hinweise für die Erstellung von Prüfungsaufgaben und deren Bewertung sowie konkrete Aufgabenbeispiele" (Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2016) sowie Lehrpläne und teilweise Schwerpunktthemen. Einheitliche Prüfungsanforderungen für die Fächer Deutsch, Englisch, Französisch und Mathematik werden ab dem Schuljahr 2016/2017 durch die vom Institut für Qualitätsentwicklung im Bildungswesen (IQB) entwickelten Bildungsstandards ersetzt (Stanat, Becker-Mrotzek, Blum, & Tesch, 2016). Mit der Erarbeitung der Prüfungsaufgaben beauftragen die obersten Schulaufsichtsbehörden oder Landesinstitute der einzelnen Bundesländer erfahrene, aktive Lehrpersonen. Aus den Vorschlägen wählt entweder die jeweilige oberste Schulaufsichtsbehörde oder eine Kommission die Abituraufgaben aus und modifiziert sie bei Bedarf. Die Kommission besteht aus Lehrpersonen, Fachreferentinnen und Fachreferenten des Landesinstituts bzw. der Schulaufsichtsbehörde, teilweise ergänzt durch Fachwissenschaftlerinnen, Fachwissenschaftler, Fachdidaktikerinnen und Fachdidaktiker. Die Schulen erhalten bundeslandweit einheitliche Aufgabenpools, aus denen meist die Kurslehrperson und/oder die Schülerinnen und Schüler die Prüfungsaufgaben nach verschiedenen Kriterien auswählen können. Diesbezüglich bestehen nicht nur Unterschiede zwischen den Bundesländern, sondern auch zwischen verschiedenen Fächern innerhalb eines Bundeslandes. (Klein et al., 2009; Kühn, 2012)

Für die *Korrektur* liegen bundesland- und fachspezifische zentrale, externe Kriterien in Form von Erwartungshorizonten oder Bewertungsmustern vor. Sie erfolgt von der jeweiligen Kurslehrperson (Erstkorrektur) und einer Fachlehrperson derselben oder einer anderen Schule (Zweitkorrektur). Bei deutlichen Abweichungen von Erst- und Zweitkorrektur entscheidet meist der/die Vorsitzende des Prüfungsausschusses (Schulleitung, Person aus der Aufsichtsbehörde oder einer anderen Schule), der/die auch eine Drittkorrektur in Auftrag geben kann. Die Klausurergebnisse und Abiturdurchschnittsnoten melden die Schulen an die zuständige Behörde. Diese teilt den Schulen ihren Stand im Vergleich zum Landesdurchschnitt mit. (Klein et al., 2009, S. 604ff.; Kühn, 2012, S. 35ff.)

Die *Abiturgesamtnote* setzt sich aus den Noten der schriftlichen und mündlichen Abiturprüfungen (insgesamt 33%) sowie den Halbjahresnoten der einzelnen Kurse der letzten zwei Jahre der Oberstufe (vier Halbjahre, insgesamt 66%) zusammen. In die Halbjahresnoten fliessen sowohl schriftliche als auch mündliche Leistungen ein. Noten aus Kursen mit erhöhtem

Anforderungsniveau/Leistungskursen haben ein grösseres Gewicht als Noten aus Kursen mit grundlegendem Anforderungsniveau/Grundkursen. Zentrale Abiturprüfungen sind insgesamt mit etwa 20% in der Gesamtnote enthalten. (Klein et al., 2009, S. 605ff.; Kühn, 2012, S. 36; Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2013)

Auch wenn zentrale Abiturprüfungen in Deutschland einen höheren *Standardisierungsgrad* aufweisen als dezentrale Abiturprüfungen, sind sie dennoch im internationalen Vergleich auf einem geringen bis mittlerem Niveau der Standardisierung angesiedelt (Klein et al., 2009, S. 615ff.; Kühn, 2012, S. 40f.; van Ackeren et al., 2015, S. 177). Zwar suggeriert der in vielen Bundesländern verwendete Begriff „Zentralabitur" Einheitlichkeit, verbirgt jedoch, dass bei den bundeslandspezifischen Ausgestaltungen die Unterschiede zwischen den Bundesländern die Gemeinsamkeiten überwiegen (gleiches gilt auch für den internationalen Kontext; Klein, 2016; Klein & van Ackeren, 2011; van Ackeren, 2007). Die Frage, ob Prüfungen zentral oder dezentral durchgeführt werden, wird vorrangig daran festgemacht, welche Akteure die Prüfungsaufgaben entwickeln. Andere Aspekte der Organisation sind dagegen von nachgeordneter Bedeutung (Kühn, 2012, S. 37ff.; van Ackeren et al., 2015, S. 175).

Kritik

Die Möglichkeit erhöhter Standardisierung und Objektivität der Korrektur durch anonyme Benotungen wird ausser in Hamburg und Sachsen-Anhalt nicht ausgeschöpft. Diese ist ebenso wie die Korrektur durch schulexternes Personal im internationalen Vergleich weiter verbreitet (Klein et al., 2009; Klein & van Ackeren, 2011; Kühn, 2012). Wenn auch nicht direkt auf das Zentralabitur bezogen, sondern auf schullaufbahnbegleitende standardisierte Tests, sieht van Ackeren den Bedarf

> einer leistungsorientierten Kultur, die aber gleichwohl Vertrauen in die professionellen Fähigkeiten von Lehrern setzt. Die professionelle Urteilskraft von Lehrkräften muss genutzt und unterstützt werden, indem ihnen nicht bei der Auswertung von Tests und Prüfungen das Mandat entzogen wird (van Ackeren, 2005, S. 28).

In eine ähnliche Richtung weist die Aussage von Good et al. (2010, S. 146), dass Lehrpersonen im Zuge von high-stakes Verfahren „abgewertet" werden, wenn einzig standardisierte Tests über den Lernfortschritt von Schülerinnen und Schülern sowie deren Versetzung in die nächste Klassenstufe entscheiden.

2.2 Zentrale Abiturprüfungen in Bremen

Organisation

Mit Verweis auf die Durchführung zentraler Abiturprüfungen in den anderen Bundesländern zur Sicherung von Standards, betont Bremen mit der Kombination aus zentralen und dezentralen Anteilen die *Ziele* der Vergleichbarkeit und der „exemplarische[n; E.M.] Vertiefungen in den Prüfungen" (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4). Weiterhin werden folgende Ziele genannt:

- Einheitliche Anforderungen für die schriftlichen Prüfungen an den Schulen des Landes Bremen werden gesichert.
- Standards und moderne Aufgabenformate bilden eine Grundlage für eine didaktische und methodische Weiterentwicklung des Unterrichts.
- Die Ergebnisse von Unterricht und Prüfungen werden vor dem Hintergrund vorgegebener Standards evaluiert.
- Die Fachlehrerinnen und Fachlehrer werden von der Erstellung der Aufgabenvorschläge für Teile der Prüfungen entlastet. (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4)

In Bremen erfolgte die *Implementation* zentraler Abiturprüfungen 2007 zunächst in allen Fächern auf Grundkursniveau und 2008 zusätzlich in den Leistungskursen der Fächer Deutsch, fortgesetzte Fremdsprachen (Englisch, Französisch, Latein, Russisch, Spanisch, Türkisch), Mathematik und Naturwissenschaften (Biologie, Chemie, Physik). In den Leistungskursen der übrigen Fächer finden nach wie vor dezentrale Abiturprüfungen statt. Damit erfolgte die Einführung des Zentralabiturs nicht nur schrittweise, sondern auch nicht flächendeckend in allen Fächern auf Leistungskursniveau. Die Schülerinnen und Schüler müssen ihre Prüfungsfächer jedoch derart wählen, dass sie in mindestens zweien zentrale Abiturprüfungen ablegen. Zudem müssen sie seit dem Schuljahr 2009/2010 eine schriftliche Abiturprüfung in Deutsch oder Mathematik absolvieren. In Abhängigkeit der von den Schülerinnen und Schülern gewählten Prüfungsfächer tragen zentrale Abiturprüfungen 17% bzw. 23% zur Gesamtnote bei. (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4; Helbig & Nikolai, 2015, S. 219ff.; Kühn, 2012, S. 27ff.)

Vor Beginn des ersten Halbjahres der Qualifikationsphase, das heisst über zwei Jahre vor den Abiturprüfungen, werden den Schulen die *Schwerpunktthemen* bekannt gegeben, die diese zu Beginn der Qualifikationsphase den Schülerinnen und Schülern nennen. Im Sinne der Transparenz beteiligen sich die Fachkonferenzen der Schulen im Vorfeld an der Diskussion der Schwerpunktthemen. Auf Basis dieser Schwerpunktthemen entwickeln vom Landesinstitut für Schule eingerichtete Fachkommissionen im Auftrag der/des für Bildung zuständigen Senatorin/Senators die *Prüfungsaufgaben*. Je nach Fach wählen entweder der Fachprüfungsausschuss oder die Schülerinnen und Schüler eine oder mehrere Aufgaben aus. (Die Senatorin für Bildung und Wissenschaft, 2010, S. 3; 2013a, S. 4; 2013b, S. 9)

Die nicht-anonymisierte *Korrektur* der Prüfungsaufgaben nehmen die jeweilige Kurslehrperson (Erstkorrektur) sowie eine Fachlehrperson derselben Schule (Zweitkorrektur) vor. Differiert die Benotung, entscheidet die/der Vorsitzende des Fachausschusses (Die Senatorin für Bildung und Wissenschaft, 2013b, S. 11f.).

> Jeder Prüfungsaufgabe wird ein Erwartungshorizont mit der Angabe von Bewertungskriterien beigegeben. In dieser Darstellung werden die unterrichtlichen Voraussetzungen benannt. In den Erwartungshorizonten werden die für die Lösung der Aufgabe vorauszusetzenden Leistungen der Prüflinge kriterienorientiert, ggf. stichwortartig auf die drei Anforderungsbereiche bezogen beschrieben. Sie enthalten auch Hinweise darauf, mit welchem Gewicht die einzelnen Anforderungsbereiche oder Aufgabenteile in die Bewertung der Gesamtleistung eingehen. (Die Senatorin für Bildung und Wissenschaft, 2010, S. 3)

Zur *Sicherung der Qualität* vergleicht die Senatorin/der Senator für Bildung und Wissenschaft die zentralen Anteile mit den Erwartungshorizonten und Korrekturhinweisen der Abiturprüfungen mit den benoteten Prüfungen der Schülerinnen und Schüler. Ausserdem wertet die Prüfungskommission mit den Fachprüfungsleitenden die schriftlichen und mündlichen Prüfungen für die Diskussion in den Fachkonferenzen der jeweiligen Schule aus. Seit August 2009 werden diese Ergebnisse für die Vorbereitung der nächsten Abiturprüfungen genutzt. (Die Senatorin für Bildung und Wissenschaft, 2013b, S. 15) Darüber hinaus erhalten die Schulen in Bremen Informationen zu ihren Ergebnissen

im Vergleich zum Landesdurchschnitt und es wird ein Landesbericht mit den Ergebnissen online zur Verfügung gestellt (Kühn, 2012, S. 37). Hinzu kommt „the fact that the members of the examination commissions in the German states are teachers in active service, and their experiences flow back into school practice in their own schools" (Maag Merki & Holmeier, 2015, S. 61). Insgesamt finden die zentralen Abiturprüfungen in Bremen im internationalen sowie nationalen Vergleich mit einem geringen Mass an Standardisierung statt (Kühn, 2012, S. 41).

Kritik

Kritik und Befürchtungen zum Zentralabitur und zu standardisierten Tests (Kapitel 2 und 2.1) lassen sich auf die zentralen Abiturprüfungen in Bremen übertragen mit der Einschränkung einiger bundeslandspezifischer Besonderheiten, wie beispielsweise die schrittweise und nicht vollumfängliche Implementation oder der Monitoringprozess unter Einbezug der Lehrpersonen. Die Kombination zentraler und dezentraler Anteile sowie die Berücksichtigung der Erfahrungen der beteiligten Akteure schwächen einige Kritikpunkte, z. B. erhöhten Stress, möglicherweise ab.

3. Theoretischer Hintergrund

Die einzelnen Beiträge (Kapitel 6; Publikationen im Anhang) decken verschiedene, mit der Implementation des Zentralabiturs einhergehende Aspekte in längerfristiger Perspektive ab. Zwar beinhaltet jeder Beitrag eine spezifische Ausrichtung mit je eigenem theoretischen Kontext bzw. Hintergrund sowie mit vielfältigen theoretischen Anknüpfungspunkten, doch haben auch alle Gemeinsamkeiten. Der dieser Arbeit zugrunde liegende weitreichendere theoretische Rahmen spannt den Bogen von der Beschreibung der Beschaffenheit des Bildungssystems als Mehrebenensystem und von Interpretationsleistungen verschiedener Akteure und Akteursgruppen (Kapitel 3.1) über unterschiedliche Perspektiven der Frage der Steuerung des Bildungssystems (Kapitel 3.2) und der (Analyse der) Handlungskoordination verschiedener mit Bildung befasster Akteure und Akteursgruppen im Mehrebenensystem (Kapitel 3.3) hin zu mittels der Implementation von Reformen und Innovationen initiierten Veränderungen (Kapitel 3.4). Der Bezug zur Schulentwicklung greift die Perspektive der Veränderungen auf und richtet den Blick auf die Ebene der Einzelschule. Dagegen verweist die Schuleffektivität auf Faktoren, die für die Leistungen des Bildungssystems, zumeist Lernergebnisse der Schülerinnen und Schüler, relevant sind (Kapitel 3.5). Der Entwurf eines eigenen Modells potentieller Veränderungen aufgrund der Implementation des Zentralabiturs (Kapitel 3.6) schliesst die Ausführungen zum theoretischen Hintergrund ab.

3.1 Mehrebenenstruktur des Bildungssystems und Rekontextualisierung

Eine theoretische Verbindung der einzelnen Beiträge stellt die von Fend (2008b) als *Mehrebenencharakter des Bildungssystems* beschriebene Struktur desselben dar. Während auf der Makroebene beispielsweise Gesetze, Regelungen und Bildungspläne der Bildungspolitik und -verwaltung zu verorten sind, befindet sich auf der Mesoebene die Einzelschule als „pädagogische Handlungseinheit" (Fend, 2008b, S. 146) mit den entsprechenden Rahmenbedingungen. Die Mikroebene lässt sich nach den Akteursgruppen Lehrpersonen sowie Schülerinnen und Schüler differenzieren. Auf Seite der Lehrpersonen sind Regelungen und Aufgaben, die ihre Rolle und ihr Handeln bestimmen, von Bedeutung. Bei den Schülerinnen und Schülern sind Regelungen und Bedingungen beispielsweise in Zusammenhang mit Prüfungen, Abschlüssen oder Zulassungen von Belang. (Fend, 2008b) Die Triade von Makro-, Meso- und

Mikroebene ist komplex, da die verschiedenen Ebenen und ihre Akteure in einem reziproken Verhältnis stehen und sich gegenseitig Umwelt sind. Fullan spricht diesbezüglich von der „Verschmelzung von Mikrokosmos und Makrokosmos" (Fullan, 1999, S. 233).

> Das Handeln auf der jeweiligen Ebene impliziert immer, dass die übergeordnete Ebene für die untergeordneten als Kontext präsent ist, aber im Rahmen der ebenenspezifischen Umweltbedingungen und Handlungsressourcen reinterpretiert und handlungspraktisch transformiert wird. Die übergeordnete Ebene bleibt also erhalten, wird aber gleichzeitig verändert. Rekontextualisierung meint deshalb Handeln im Rahmen von Ordnungen des Zusammenhandelns angesichts gegebener Umwelten, vermittelt durch die Selbstreferenz, die Interessen und Ressourcen der Handelnden. (Fend, 2008a, S. 181)

Rekontextualisierungsprozesse ereignen sich als Interpretation und Umsetzung übergeordneter Rahmenbedingungen (Handlungskontext 1) sowie als Handlungs- und Rahmenbedingungen der jeweiligen Ebene mit je eigener Logik (Handlungskontext 2; Fend, 2008b, S. 34). Dabei sind die „Interdependenzen der einzelnen Ebenen aufschlussreich […] für das Verstehen der Handlungen der Akteure und der Leistungen der Systeme" (Altrichter, 2015, S. 37f.). Die Vorgaben sind selbst bei individueller Umsetzung handlungsleitend und bindend. Sie werden durch reflexive Prozesse an den entsprechenden Kontext und dessen Handlungsaufgaben adaptiert und verändert, wobei die jeweiligen professionellen Einstellungen und Deutungsmuster der Akteure eine Rolle spielen. Somit ist nicht nur eine top-down-Richtung denkbar, sondern auch eine bottom-up (Fend, 2008b, S. 27), wobei Brüsemeister kritisiert, dass die zweite Perspektive, ebenso wie „seitwärtige Ebenen", keine Beachtung finden (Brüsemeister, 2010, S. 14). Zudem sind die spezifischen Handlungslogiken der einzelnen Akteure, nicht nur der verschiedenen Ebenen, zu bedenken (Kussau & Brüsemeister, 2007, S. 33; Reusser & Halbheer, 2008, S. 128f.). Insgesamt ist davon auszugehen, dass auf allen Ebenen des Bildungssystems Rekontextualisierungsprozesse mit je spezifischen Referenzrahmen stattfinden: Auf der Makroebene steht die Frage im Zentrum, welches Weltbild und welche Bereiche das Bildungssystem in welcher Intensität vermitteln soll (institutionelles Programm). Diese Frage wird auf der Mesoebene mittels der Auswahl und Organisation von Themen und Wissensbeständen, das heisst Bildungsinhalten, im Spannungsfeld von einerseits politischen und administrativen und andererseits (fach-)didaktischen Rahmenbedingungen

konkretisiert (kulturelles Programm). Diese Bildungsinhalte adaptieren auf der Mikroebene Lehrpersonen und setzen sie in Unterrichtseinheiten um. Hierbei berücksichtigen sie erstens die Lernbedingungen und -möglichkeiten der Schülerinnen und Schüler (primäre Rekontextualisierung), zweitens Aspekte institutioneller Rahmenvorgaben wie Prüfbarkeit und Prüfung von Inhalten, Sicherung von Standards (sekundäre Rekontextualisierung) und drittens Handlungsbedingungen eines „komplexen sozialen Erwartungszusammenhang[s; E.M.] von Elternhaus, Gemeinde und Öffentlichkeit" (sekundäre Rekontextualisierung; Fend, 2008b, S. 239ff., 331; auch Berry & Adamson, 2011, S. 9f.). Primäre und sekundäre Rekontextualisierung stehen zudem in einer interdependenten Beziehung zur Innenwelt der Lehrperson. Die Schülerinnen und Schüler als Nutzerinnen und Nutzer bewegen sich ihrerseits im Spannungsfeld erstens des schulischen Angebots, zweitens der eigenen Erwartungen, Bedürfnisse und Ressourcen und drittens den Erwartungen wichtiger Bezugspersonen (Fend, 2008b, S. 29ff.). Diese Innenwelt der Schülerinnen und Schüler verweist auf die Nähe zur Definition von Sozialisation als „produktive Verarbeitung der inneren und äußeren Realität" von Hurrelmann (2006, S. 11). In der Sozialisationsinstanz Schule verarbeiten Kinder und Jugendliche die durch Vorgaben und Rekontextualisierungsprozesse der verschiedenen Ebenen und Akteure gestaltete äussere Umwelt bzw. Realität in Abhängigkeit ihrer inneren Realität, nach Fend insbesondere ihrer realen und wahrgenommenen Bedürfnisse, Wünsche, Motivationen, Lernkapazitäten, Reaktionen, Ziele, beliefs, Freuden und Ängste (Fend, 2008b, S. 33).

Für die Implementation *zentraler Abiturprüfungen* bedeutet dies, dass die auf der Makroebene formulierten institutionellen Regeln und Vorgaben (z. B. für Prüfungen) auf der Mesoebene an einzelschulspezifische Bedingungen adaptiert, Handlungsweisen und Abläufe bei Bedarf modifiziert sowie Ressourcen zur Vorbereitung und Durchführung zentraler Abiturprüfungen bereitgestellt werden. Lehrpersonen als Akteure der Mikroebene sind bestrebt, ein auf die Bedingungen der Klasse und der Schülerinnen und Schüler individuell abgestimmtes Angebot zur Passung zwischen den Vorgaben (z. B. Schwerpunktthemen, Prüfungsanforderungen und -formate) und ihrem Unterricht bereitzustellen. Dieses nutzen die Schülerinnen und Schüler ebenfalls als Akteure der Mikroebene entsprechend ihren individuellen Bedingungen. Denkbar wäre, dass sich durch die Einführung zentraler Abituraufgaben bei den Lehrpersonen das Verhältnis der Gewichtung von „kulturellem Programm" und Nutzungsbedingungen der Schülerinnen und Schüler verschiebt und sie ihren Unterricht beispielsweise stärker an den institutionellen Vorgaben ausrichten als noch unter den Bedingungen dezentraler Prüfungsorganisation.

Die Synchronisierungsqualität von Angebot und Nutzung zeigt sich in den Lernergebnissen (Fend, 2008b, S. 22), unter anderem operationalisiert über die anhand der zentralen Korrekturkriterien bewerteten zentralen Abiturprüfungen (Kapitel 3.6).

Rekontextualisierungsprozesse der einzelnen Schulen als pädagogische Handlungseinheiten (Fend, 2008b) und ihrer Akteure zeigen, dass zentrale Vorgaben und Impulse nicht direkt, unverändert und einheitlich umgesetzt werden. „Wie auch immer sich die Lösung der Bearbeitung von fremd und eigen induzierten Vorgaben gestaltet, herauszustreichen ist die je lokale Lösung, die die Gruppen für ihre Schule gemäß deren Eigenlogik entwickeln" (Baum, 2014, S. 248).

Abs, Brüsemeister, Schemmann und Wissinger sprechen sogar davon, „dass es sehr viele und sehr unterschiedliche Mehrebenensysteme im Bildungsbereich gibt" (Abs et al., 2015, S. 8). Insofern stellt sich die Frage, wie Bildungssysteme trotz erheblicher „Gestaltungsfreiheit der einzelnen Schule" und der „geringen Möglichkeiten der bildungspolitischen Akteure […], mit Kontrolle oder Sanktionen auf abweichendes Verhalten der schulischen Akteure zu reagieren" (Baum, 2014, S. 248; auch Baumert, 2016, S. 222; van Ackeren et al., 2015, S. 105f.; van Ackeren et al., 2011, S. 174), gesteuert werden können und welche Rolle zentrale Abiturprüfungen dabei spielen.

3.2 (Neue) Steuerung und Steuerungsfunktion zentraler Abiturprüfungen

Die Frage nach der bestmöglichen Steuerung des Bildungswesens wird im deutschsprachigen Raum seit Anfang der 1990er Jahre intensiv diskutiert – in einer „Phase der Schulmodernisierung" (Altrichter, Rürup, & Schuchart, 2016b, S. 107). Sie ist jedoch keineswegs neu, wie Herrmann (2009) mit Verweis auf die Entwicklung seit dem „Allgemeinen Landrecht für die preußischen Staaten 1794" zeigt. Das Spannungsfeld zwischen zentralstaatlichen Regelungen und regionalen bzw. lokalen Rahmenbedingungen besteht bereits seit langem. Schulische Steuerung bezieht sich auf die Bereiche Kontext, Input, Prozess und Wirkungen bzw. Output und Outcome und hat zum Ziel, die „Beliebigkeit von Folgehandlungen" (van Ackeren et al., 2015, S. 118) zu minimieren. Der vielbeschriebene Paradigmenwechsel von der Inputsteuerung zur Outputsteuerung, auch Outcomesteuerung, Wirkungssteuerung oder Neue

Steuerung, am Ende des 20. Jahrhunderts[4] basiert auf der „Annahme, die klassischen bürokratisch-regulativen Steuerungsverfahren seien an die Grenzen ihrer Leistungsfähigkeit gestoßen" (Herrmann, 2009, S. 60; auch Maritzen, 2008, S. 109f.; Parreira do Amaral, 2012; van Ackeren et al., 2015, S. 115ff.) und geht mit Schlagworten wie Autonomie und Selbstständigkeit für die Einzelschulen, Kompetenzorientierung, Evaluation, Monitoring, Dezentralisierung, Deregulierung, Accountability, Evidenzbasierung uvm. einher (Abs et al., 2015; Altrichter & Maag Merki, 2010, 2016b; Böttcher, 2012; Fend, 2008b; Herrmann, 2009; Hornberg & Parreira do Amaral, 2012; Klemm, 2005; LISUM Deutschland, bm:ukk Österreich, & EDK Schweiz, 2008; van Ackeren & Klemm, 2011; van Ackeren et al., 2015; van Ackeren et al., 2011; Woessmann, Luedemann, Schuetz, & West, 2009).[5] Mit der Verschiebung des Fokus auf „schulische Arbeitsresultate als Qualitäts- und Steuerungsdimension" (van Ackeren & Bellenberg, 2004, S. 125) halten Output- und Wettbewerbsorientierung Einzug in die Schulen. Die „Vier E" – Effektivität (Grad der Zielerreichung), Effizienz (Kosten-Nutzen, Ressourcenverbrauch), Evidenz (Nachweis der Zweckerreichung), Erfolgsanreize – bestimmen die Arbeit(sbedingungen) verschiedener pädagogischer Akteure und Akteursgruppen (Böttcher, 2005, S. 103f.; Böttcher, Dicke, & Ziegler, 2012, S. 7f.).

Im Gegensatz zu Zeiten der Inputsteuerung verfügen die Einzelschulen nun über einen grösseren Handlungsspielraum, tragen mehr Verantwortung für ihr Handeln, ihre Ressourcen und ihre Leistungen, müssen jedoch im Gegenzug darüber stärker Rechenschaft ablegen (accountability). Parreira do Amaral spricht von einem „internationalen accountability turn" (Parreira do Amaral, 2012, S. 83), Mansell mit Blick auf England von „hyper-accountability" (Mansell, 2011, S. 294f.). Die Einzelschulen befinden sich damit in einem Spannungsfeld von Dezentralisierung oder Deregulierung auf der einen Seite und (Re-) Zentralisierung auf der anderen Seite.

> Der Alltag schulischer Akteure ist häufig durch ein scheinbar widersprüchliches Nebeneinander von bürokratischer (Über-)Regulierung und gleichzeitig existierenden Freiräumen bestimmt, denn Schule ist durch ein Spannungsfeld gekennzeichnet, das sich zwischen loser und fester Kopplung bewegt (Baum, 2014, S. 247; auch van Ackeren & Bellenberg, 2004, S. 127).

[4] Zur Kritik an dem Paradigmenwechsel und der damit zusammenhängenden Reform des deutschen Schulsystems siehe beispielsweise Böttcher (2012), Dreßler (2016) oder Zlatkin-Troitschanskaia (2007).

[5] Zu (nicht-)intendierten Folgen für das Vertrauen, etwa der Wandel des Vertrauens in Institutionen, Bildung und Professionalität der mit Bildung beschäftigten Akteure hin zu einem Vertrauen in Kontrolle, Rechenschaftslegung, Instrumente und Zahlen sowie damit einhergehendes Misstrauen, siehe Bormann (2012; auch Penninckx, Quintelier, Vanhoof, De Maeyer, & Van Petegem, 2017).

Es handelt sich demnach weniger um dichotome entweder-oder-Kategorien als vielmehr um graduelle Niveauunterschiede auf einem Kontinuum. Als Beispiel sei hier auf die auf die dezentralen Elemente beim „Zentralabitur" (Kapitel 2.1, 2.2) oder die unterschiedlichen Abstufungen und Bereiche von Schulautonomie verwiesen (Altrichter et al., 2016b; Dreßler, 2016; Rürup & Heinrich, 2007; Smyth, 2011; Woessmann et al., 2009; Zlatkin-Troitschanskaia, Förster, & Preuße, 2012).

Die Steuerungselemente der deutschsprachigen Länder gleichen sich auf den ersten Blick. Der zweite Blick verrät jedoch Differenzen zwischen den Ländern, sodass Altrichter und Maag Merki (2016a, S. 22f.) dafür plädieren, von mehreren Steuerungsmodellen statt von einem auszugehen. Ihre Veränderungen bedeuten insgesamt eher eine Ergänzung als eine Ablösung bestehender Steuerungselemente (Altrichter & Maag Merki, 2016a, S. 23; Kussau & Brüsemeister, 2007, S. 42f.; von Recum, 2003, S. 106f.; Zlatkin-Troitschanskaia, 2007, S. 80f.). Der Steuerungsdiskurs der letzten Jahre steht unter der „Leitfrage":

> Wie kann die Steuerungsstruktur des Schulwesens (die Art und Weise, wie seine Ordnung und seine Leistung zustande kommen und sich weiterentwickeln) rasch und zielgerichtet so verändert werden, dass qualitätsvolle Ergebnisse – und bessere Ergebnisse als bisher – ökonomisch erbracht werden können (Altrichter & Maag Merki, 2016a, S. 3)?

Voraussetzung für die Beantwortung dieser Frage mit ihrem Fokus auf Output und Outcome ist entsprechendes Steuerungswissen, welches mittels *Systemmonitoring* gewonnen wird. Funktionen des Systemmonitorings sind

> die Beobachtung, Analyse und Darstellung wesentlicher Aspekte eines Systems, verbunden mit der Funktion der Systemkontrolle einschließlich der Angleichung von Leistungsmaßstäben *(Benchmarks)* sowie die Funktion, „Steuerungswissen" zu generieren bzw. zu erweitern und „Steuerungshandeln" begründbarer und zielgerichteter zu gestalten (Döbert, 2008, S. 74f.; Hervorhebung im Original).

Das *Bildungsmonitoring* „als Oberbegriff für die vielfältigen Datenerhebungen im Bildungswesen" (Rürup, Fuchs, & Weishaupt, 2016, S. 413) soll Informationen über sämtliche Bildungsinstitutionen als Grundlage für Diskussionen und Entscheidungen liefern. Als Grundfunktionen lassen sich – ähnlich wie die Funktionen des Systemmonitorings – Akkreditierung bzw. Zertifizierung, Rechenschaftslegung sowie

Diagnostik für systemisches Lernen differenzieren (Maritzen, 2008, S. 110f.). In Deutschland stützt sich das Bildungsmonitoring auf

- internationale Schulleistungsuntersuchungen,
- zentrale Überprüfung des Erreichens der Bildungsstandards in einem Ländervergleich (in der 4., 9. und 10. Klasse),
- Vergleichsarbeiten in Anbindung an die Bildungsstandards zur landesweiten Überprüfung der Leistungsfähigkeit einzelner Schulen,
- gemeinsame Bildungsberichterstattung von Bund und Ländern (Döbert, 2008, S. 75; ausführlicher Maritzen, 2008, S. 112ff.; 2011, S. 121f.).

Zum Teil werden zusätzlich Schulinspektionen zum Bildungsmonitoring gezählt (Altrichter, Brüsemeister, & Wissinger, 2007). Es umfasst in Deutschland demzufolge internationale, nationale sowie bundeslandspezifische Elemente und hat grossen Einfluss auf das Zusammenspiel verschiedener Akteure auf unterschiedlichen Ebenen des Bildungssystems. Der vorrangige Fokus liegt dabei auf der Makroebene und seltener auf der Mesoebene.

> […] the relationships between the different levels of the system (state, region, school, class), between the stakeholders (parents, teachers, principals, superintendents) and the instruments and procedures (normative conditional programming, empirical monitoring of the processes and effects of education) are being thoroughly reconfigured (Maritzen, 2011, S. 119).

Die einzelnen Elemente des Bildungsmonitorings in Deutschland weisen eine hohe Ähnlichkeit zu den Komponenten des „standards-based accountability system" im Zuge des *No Child Left Behind*-Aktes in den USA auf, welches sich auf vier wesentliche Elemente stützt:

- Content standards that set out the knowledge and skills children are expected to learn.
- Tests or assessments to measure those content standards.
- Student performance standards that define proficient performance in terms of the official assessments.
- Rewards provided to students or schools that meet or exceed the standards and punishments or remediation activities for those that do not. (Firestone & Schorr, 2004, S. 5)

Einzig der vierte Punkt der high-stakes Sanktionen findet in Deutschland keine Anwendung.

Als *Instrumente der Neuen Steuerung* gelten unter anderem zentrale Abschluss- und Abiturprüfungen, Vergleichsarbeiten, Lernstandserhebungen zur Überprüfung von Bildungsstandards sowie externe Schulinspektionen (Kühn, 2010; Maag Merki, 2016; Maag Merki & Emmerich, 2011; van Ackeren & Bellenberg, 2004). Damit sind sie zum Teil deckungsgleich mit den Verfahren des Bildungsmonitorings, zum Teil stellen sie jedoch eine Ergänzung dar, um die Perspektive der Einzelschule zu stärken (Mesoebene). Ziel ist die Überprüfung der Arbeit der Schulen, vorrangig gemessen als Leistungen der Schülerinnen und Schüler, sowie die Sicherung und ggf. Steigerung der Qualität von Schule und Unterricht. Die Schulen müssen demnach im Zuge ihrer, durch erweiterte Autonomie gesteigerten, Handlungsmöglichkeiten die von der Makroebene zentral vorgegebenen und kontrollierten Standards einzelschulspezifisch erfüllen. Wie sie die definierten Leistungen erbringen, ist den Schulen überlassen (Böttcher, 2012, S. 35ff.: „Fehlsteuerung"; auch Hahn, 2014, S. 323f.). Allerdings lässt sich die „für das New Public Management charakteristische Aufteilung von Zielen und Mitteln auf Prinzipale und Agenten" (Bellmann, 2016, S. 20) nicht ohne Schwierigkeiten auf das Bildungssystem übertragen, da zum einen Ziele und Mittel interdependent sind (Bellmann, 2016, S. 14) und zum anderen die Beziehungen zwischen den Akteuren zahlreicher und komplexer als im wirtschaftlichen Bereich sind.

> In the education process, a network of principal–agent relationships exists that entail conflicts between the interests of different groups—mainly students, parents, teachers, heads of school, administrators, and the government—and serious problems of monitoring due to informational advantages of self-interested agents. This can create adverse incentives and leeway for the agents to act opportunistically, leading to an inefficient use of given resources and to misallocations of resources across different uses. By determining decision-making rules and incentives, the institutional structure of the schooling system can thus influence the quality of the education that is ultimately produced. (Bishop & Wößmann, 2004, S. 18)

Steuerungen im Bildungssystem müssen den Handlungs- und Entscheidungsrahmen der wechselseitig verwobenen Akteure und Akteursgruppen mit je eigenen Interessen so weit wie nötig und so eng wie möglich gestalten, um zu grosse Freiräume für abweichendes Verhalten einzugrenzen. Dabei kann die „gegenseitige Nicht-Einsehbarkeit der unterschiedlichen Handlungsebenen und -kontexte" (Lange, Rahn, Seitter, & Körzel, 2009, S. 10) eine zielgerichtete Steuerung erschweren.

An der Neuen Steuerung kritisiert Herrmann (2009, S. 60; ähnlich Bellmann, 2016, S. 14; Böttcher, 2012), dass diese für sich in Anspruch nimmt, Grundbedingung für eine Verbesserung der Qualität von Unterricht und Lernergebnissen zu sein, obwohl empirische Belege für die Bedeutung unterschiedlicher Steuerungsformen und deren Auswirkungen auf die Qualität von Lehren und Lernen fehlen.

Dem monierten Informationsdefizit begegnet eine umfassende Forschung zu vielfältigen Effekten zentraler und dezentraler Steuerung sowie zu differenziellen Auswirkungen des Wechsels von dezentralen zu zentralen Steuerungselementen (Kapitel 4).

Steuerungsfunktion zentraler Abschlussprüfungen

Das *standards-based accountability system* in den USA (Firestone & Schorr, 2004, S. 5) geht davon aus, dass die Definition und Transparenz von Lernzielen mittels Standards in Kombination mit der Erhebung des Lernfortschritts das Lernen verbessert. Dabei wirken die Standards einerseits auf das externe System der Leistungsüberprüfungen, andererseits auf die Handlungen in der Schule und im Unterricht (Herman, 2005, S. 1ff.; Schraw, 2010, S. 73), was auch kritisch gesehen wird:

> Die Steuerungsgelüste sickern gleichsam von oben nach unten: von der Bildungsadministration zur Leitung der einzelnen Schule, von der Schulleitung zu den Lehrkräften und von diesen zu den Schülerinnen und Schülern, deren Lernverhalten bis ins Detail als kontrollierbar erscheint, *und zwar von ganz oben* (Herzog, 2016, S. 132; Hervorhebung im Original; auch Mansell, 2011, S. 293).

Die Triade aus Standards, Prüfungen und Praktiken in Schule und Unterricht skizzieren Barnes, Clarke und Stephens ähnlich, positionieren mit Blick auf die Gegebenheiten in Australien jedoch die Prüfungen anders: „mandated assessment mediates between the expectations of the system and their embodiment in classroom practice" (Barnes, Clarke, & Stephens, 2000, S. 626). Vor dem Hintergrund des *standards-based accountability system* in den USA berücksichtigt die *standards-based accountability theory of action* (Hamilton, Stecher, Russell, Marsh, & Miles, 2008) ebenfalls verschiedene Ebenen sowie Rekontextualisierungsprozesse (Kapitel 3.1). Demnach beeinflussen Bildungsreformen die

Handlungen von Lehrpersonen direkt und indirekt über die Reaktionen der Schule. Letztere steht als „unit of accountability" im Fokus des *standards-based accountability system*, welche das gesamte Bildungssystem hin zu einer höheren Verbindlichkeit und Verbindung von Zielen, Handlungen und Outcome reformieren soll. Dafür sind jedoch Anstrengungen und Ressourcen aller Ebenen des Bildungssystems zu koordinieren (Hamilton et al., 2008, S. 32ff.; Herman, 2005, S. 2). Ausgangspunkt der *standards-based accountability theory of action* sind landesweite Standards, Instrumente zu deren Überprüfung sowie Anreize, Informationen und Unterstützungsleistungen. Diese Komponenten bedingen die Implementation der Standards in den *districts* und in den Einzelschulen, welche sich wiederum auf das Handeln der Lehrpersonen im Unterricht auswirkt und sich schliesslich in den Lern- und Testergebnissen der Schülerinnen und Schüler niederschlägt. Dabei beeinflussen Meinungen, Haltungen, sense-making und Emotionen der jeweiligen Akteure wie auch hinderliche oder förderliche (organisatorische) Faktoren die Reaktionen der *districts*, der Einzelschulen sowie das Handeln im Unterricht. Hürden sind beispielsweise föderale Strukturen des Bildungssystems, ein hohes Mass an Autonomie seitens der Lehrpersonen sowie deren traditionelle Übereinkunft, sich nicht in die Belange von Kolleginnen und Kollegen einzumischen (Hamilton et al., 2008, S. 34f.; Herman, 2005, S. 1f.; Kelchtermans, 2005, S. 1003). Die Lern- und Testergebnisse der Schülerinnen und Schüler stellen nicht nur den (vorläufigen) Endpunkt eines Prozesses dar, sondern sind zugleich Grundlage für Anreiz-, Informations- und Unterstützungssysteme. Somit führen Tests am Schuljahresende mittels eines „feedback loops" zu positiven oder negativen Konsequenzen im folgenden Schuljahr. Ziel ist die Abkehr von ineffizienten Praktiken und der Aufbau bzw. die Verstärkung funktionierender Handlungsweisen (Hamilton et al., 2008, S. 34f.; Herman, 2005, S. 1f.).

Diese beschriebene Ableitung von Handlungen bzw. deren notwendigen Modifizierungen aus Rückmeldungen von Ergebnissen vergangener Lernstandserhebungen oder zentraler Abschlussprüfungen greift Maag Merki (2016, S. 159f.) in ihrem Wirkungsmodell auf. Vergleiche der rückgemeldeten Ist-Ergebnisse mit den Soll-Ergebnissen (Standards) sollen zur Stärkung funktionierender Prozesse bzw. zur Reflexion von zu überarbeitenden Handlungen und Prozessen sowie zur Ableitung von Massnahmen der Schulentwicklung führen. Ziel ist es, durch das Setzen und Überprüfen von Standards die Leistungen der Schülerinnen und Schüler zu steigern sowie die Massstäbe von Benotungen zu vereinheitlichen (Maag Merki, 2016, S. 160).

Diverse Faktoren auf Seiten der verschiedenen Ebenen, Akteursgruppen und deren Interaktion beeinflussen derartige Prozesse, sodass nicht von einem Muster der Verarbeitung, sondern von vielfältigen Erscheinungsformen auszugehen ist. Auf der Mikroebene spielen bei Lehrpersonen insbesondere deren professionelle Kompetenz, Handlungsstrategien im Feedbackprozess sowie Erfahrungen mit Lernstandserhebungen oder zentralen Abschlussprüfungen eine Rolle. Letztere sind auch bei Schülerinnen und Schülern ebenso wie individuelle und kollektive Lernumwelten und -voraussetzungen von Bedeutung (Maag Merki, 2012b, S. 20f.). Zentralen Abschlussprüfungen wird dabei aufgrund ihres externen Referenzmassstabes zur Beurteilung der Leistung ein höherer „Wert" zugeschrieben als dezentralen Prüfungen, was mit einer geringeren ablehnenden Haltung dem Lernen gegenüber unter den Peers und einer stärkeren Überwachung des Lernprozesses seitens der Schülerinnen, Schüler, Eltern, Lehrpersonen und Schulen einhergehen sollte (Bishop & Wößmann, 2004; Piopiunik, Schwerdt, & Wößmann, 2014; Wößmann, 2003). Auf der Mesoebene erzeugen kollektive Handlungen von Lehrpersonen und Schulleitung(en), deren Überzeugungen sowie Erfahrungen mit Schulentwicklung Varianz zwischen den einzelnen Schulen im Rahmen von auf der Makroebene installierten Monitoring- und Rechenschaftssystemen sowie den Bedingungen des Kontextes (Maag Merki, 2012b, S. 19ff.; 2016, S. 160).

Die einzelnen Ebenen, Akteure sowie inner- und ausserschulische gesellschaftliche und (bildungs-) politische Faktoren sind vielfältig miteinander verbunden und wirken sich als unterschiedliches „Gesamtpaket" auf schulische Prozesse aus. Insofern kann davon ausgegangen werden, dass mit der Implementation des Zentralabiturs sowohl direkte Wirkungen auf das Handeln und Erleben der Akteure und Akteursgruppen, insbesondere der Schülerinnen, Schüler und Lehrpersonen, als auch verzögerte Effekte einhergehen. Diese beiden Arten von Effekten können parallel wirken oder sich ergänzen (Maag Merki, 2012b, S. 18f.). Hinzu kommen Differenzierungen nach der Wirkungsrichtung (positiv vs. negativ) sowie nach der Stabilität (kurzfristig vs. längerfristig). Maag Merki (2014, S. 63f.) stellt insgesamt fünf verschiedene Möglichkeiten von Auswirkungen der Einführung zentraler Abiturprüfungen heraus (Abbildung 1). Auch der Fall, dass sich keine Veränderungen zeigen (Nummer 1), wird berücksichtigt.

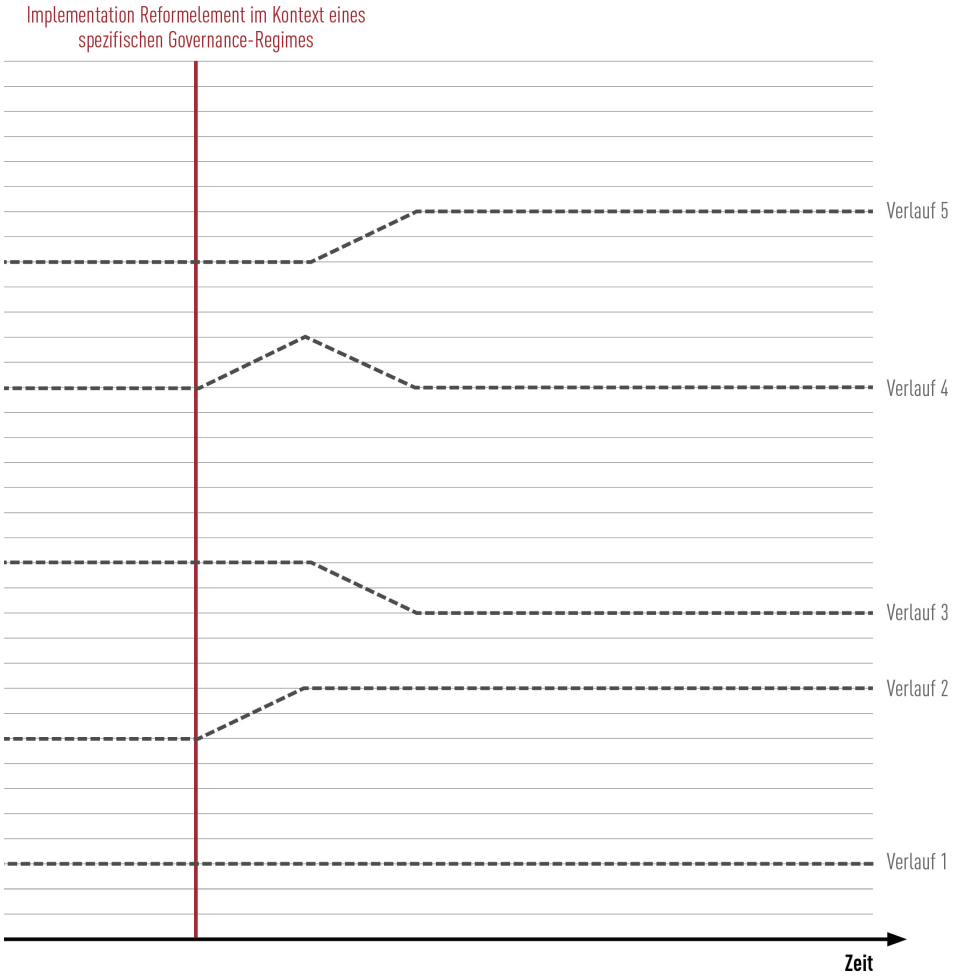


Abbildung 1: Modell der möglichen Entwicklungen/Wirkungen in Abhängigkeit der Zeit (Maag Merki, 2014, S. 63); Eigene Darstellung

Kritisch ist anzumerken, dass bei diesem Modell die Verläufe 4 und 5 auch umgekehrt denkbar sind. So kann es nach der Implementation zentraler Abiturprüfungen zunächst zu einer Reduktion einzelner Dimensionen kommen, beispielsweise der (gefühlten) Freiheiten bei der Unterrichtsgestaltung. Mit der Zeit wird jedoch das Niveau vor der Reform erreicht. Dass sich kurzfristige Variationen längerfristig abschwächen bzw. verstärken und damit das Niveau im Vergleich zur Zeit vor der Reform höher bzw. niedriger liegt (Modifikation von Verlauf 4), ist ebenfalls nicht berücksichtigt. Zudem erscheint der vor der Reform betrachtete Zeitraum verhältnismässig kurz. Die Entscheidung für die Implementation zentraler Abiturprüfungen veränderte bereits die Ausgangslage. Die Schwerpunktthemen der zentralen Abiturprüfungen in Bremen werden mehr als zwei Jahre vor der Durchführung bekannt gegeben. Damit waren zentrale Abiturprüfungen für die Grundkurse spätestens 2005 im Fokus des Interesses und der

Diskussion der verschiedenen Akteursgruppen. Eine Erhebung der dezentralen Situation hätte demnach vor der Entscheidung für zentrale Abiturprüfungen erfolgen müssen. Dies ist in der Praxis allerdings nur schwerlich bis gar nicht umzusetzen.

3.3 Educational Governance

Ungeachtet der Komplexität und der Hürden zentraler Steuerung bescheinigen Abs et al. bildungspolitischen Akteuren verschiedener Ebenen „eine Rezentralisierung und ein[en; E.M.] erstarkte[n; E.M.] Steuerungsoptimismus mittels staatlicher Intervention" (Abs et al., 2015, S. 7; auch Zlatkin-Troitschanskaia, 2007, S. 80).

Diese kann der Staat jedoch nicht alleine umsetzen. Vielmehr müssen staatliche und nichtstaatliche Akteure Prozesse aushandeln. Darauf verweist der Begriff *Governance*, der seinen Ursprung in den Politikwissenschaften und dort die Konzepte „Steuerung" und „Planung" abgelöst hat (Benz, Lütz, Schimank, & Simonis, 2007, S. 12). Governance wird verwendet als deskriptiver Begriff für „Regulierungs- und Steuerungsverhältnisse in Mehrebenensystemen", als normativer Begriff (aktuelle Modelle von Steuerung und Regulierung) oder praktisches Konzept, das heisst Analyse- und Forschungsperspektive (Altrichter, 2015, S. 33f.; Benz et al., 2007, S. 13f.). Governance kann, meist normativ, „verschiedene Formen der absichtsvollen Regelung kollektiver Sachverhalte" (Maynitz, 2009, S. 8) oder im weiteren Sinn soziale Handlungskoordinationen umfassen (Altrichter, 2015, S. 26f.). Ziel ist es, „Strukturen, Mechanismen und Wirkungen der Bewältigung von Interdependenz zwischen individuellen, kollektiven oder korporativen Akteuren zu beleuchten" (Benz et al., 2007, S. 18). Das Handeln der Akteure bestimmen dabei nicht nur Governance-Strukturen mit „Anreize[n; E.M.] und Restriktionen" (Benz et al., 2007, S. 19), die sowohl „Kontext des Handelns als auch Gegenstand von formellen und informellen Gestaltungsbemühen der Akteure" (Altrichter, 2015, S. 28) sind, sondern auch verschiedene Akteurskonstellationen und Organisationen auf unterschiedlichen Ebenen. Handlungen werden wechselseitig adaptiert und leisten einen je spezifischen Beitrag zu gemeinsamen, abgestimmten Ergebnissen (Abs et al., 2015, S. 12; Altrichter, 2015, S. 28; Parreira do Amaral, 2012, S. 75). Benz verwendet hierfür den Begriff „Funktionslogik". Er versteht darunter „ein kollektives Handeln und Politikergebnis, das durch Regeln (Institutionen) und

relativ dauerhafte Akteurs- und Interaktionskonstellationen bewirkt wird" (Benz, 2009, S. 50). Die Komplexität der Verbindungen unterschiedlicher Akteure und Akteurskonstellationen erschwert dabei eine deutliche Differenzierung „zwischen Steuerungssubjekten und -objekten" (Altrichter, 2015, S. 28; auch Koch, 2009, S. 118). Regeln und Rahmenbedingungen des Handelns können im Widerspruch zueinanderstehen, was nicht nur zu Koordinationsschwierigkeiten und Handlungsunfähigkeit, sondern auch zu Handlungsfreiheiten führen kann, Entscheidungen für oder gegen eine bestimmte Richtung zu treffen. Zuständigkeiten und Rechenschaftspflichten müssen festgelegt sein (Benz, 2009, S. 51; ähnlich Altrichter, 2015; Parreira do Amaral, 2012, S. 75).

Der Begriff Governance im Mehrebenensystem (*multilevel governance*) betont zum einen die Interdependenzen des Handelns auf verschiedenen Ebenen, die eine Koordination der Handlungen erfordern, da Kompetenzen und Ressourcen auf mehrere Ebenen verteilt sind (vertikale Beziehungen). Zum anderen handeln auf jeder Ebene mehrere, unterschiedliche Akteure und Akteurskonstellationen mit je eigenen Handlungslogiken, die ebenfalls in Beziehung stehen und ihre Handlungen koordinieren müssen (horizontale Beziehungen) (Benz, 2009, S. 15; 17). Hier zeigen sich Parallelen zur Konzeption des Bildungssystems als Mehrebenensystem von Fend (2008a; 2008b; Kapitel 3.1): Handeln wird über mehrere Ebenen hinweg durch Interpretation und Umsetzung von Vorgaben koordiniert (vertikale Beziehungen) und an die ebenenspezifischen Rahmenbedingungen, Handlungsspielräume und Akteurskonstellationen adaptiert (horizontale Beziehungen). Rekontextualisierung umfasst dabei Prozesse der Handlungskoordination verschiedener Akteure auf unterschiedlicher wie auch auf gleicher Ebene, die sich gegenseitig als Referenzsystem dienen. Insofern trifft das folgende Zitat, ursprünglich für den Bereich der Politik gedacht, gleichermassen auf den Bildungsbereich zu.

> Politische Mehrebenensysteme werden also weder von einem Zentrum aus regiert, noch werden öffentliche Aufgaben nach Ebenen getrennt innerhalb von staatlichen Gebietseinheiten erfüllt. Regieren beruht auf dem Zusammenwirken von inter- und intragouvernementalen Strukturen und Prozessen. (Benz, 2009, S. 15)

Als Analyseperspektive bietet Governance den Vorteil einer inter- und intradisziplinären Anschlussfähigkeit und die Möglichkeit, verschiedene Theorien und Methoden zu verbinden (Benz et al., 2007, S. 17;

Gördel, 2015). Sie wird in verschiedenen wissenschaftlichen Bereichen und Disziplinen angewendet. So befasst sich beispielsweise insbesondere die Politikwissenschaft, aber auch die Rechts- und Verwaltungswissenschaft, mit diversen nationalen, internationalen, europäischen und globalen Politikfeldern sowie Fragen des Föderalismus, des Wettbewerbs, von Finanzen und Finanzmärkten (Benz, 2009; Maynitz, 2009; Speyer, 2006), von Umweltpolitik (Biermann & Pattberg, 2004; Jauß & Stark, 2004; Köck, 2006; Roßnagel & von Wangenheim, 2010) oder mit Governance-Reformen in Staat, Unternehmen und Zivilgesellschaft (Klenk & Nullmeier, 2004). Von Seiten der Soziologie interessiert der Zusammenhang von Governance mit unterschiedlichen Aspekten der gesellschaftlichen Integration (Lange & Schimank, 2004), wie etwa das Gesundheitssystem (Bandelow, 2004; Bode, 2010; Hänlein & Schroeder, 2010) oder die Rolle von Massenmedien (Hofmann, 2006; Jarren & Donges, 2004) und der Wissenschaft (Braun, 2004; Kehm & Fuchs, 2010).

Die Übertragbarkeit des Konzeptes der Governance auf den Bildungsbereich erfolgt unter dem Begriff *Educational Governance*. Nach Altrichter ist es ein

> Forschungsansatz einer interdisziplinären Bildungsforschung, der
> - das Zustandekommen, die Aufrechterhaltung und die Transformation sozialer Ordnung und sozialer Leistungen im Bildungswesen
> - unter der Perspektive der Handlungskoordination
> - zwischen verschiedenen Akteuren
> - in komplexen Mehrebenensystemen untersucht (Altrichter, 2015, S. 35).

Einzig der erste Punkt stellt einen expliziten Bezug zum Bildungsbereich her, die übrigen Aspekte gelten für sämtliche Bereiche, auf die das Konzept der Governance angewendet wird. Das spezifische der Educational Governance liegt auf den ersten Blick vorrangig im Analysefokus auf Bildung. In einem weiteren Schritt werden Unterschiede zu anderen gesellschaftlichen Bereichen deutlich, beispielsweise durch je eigene Akteure und Akteurskonstellationen. Darüber hinaus bietet das Konzept der Governance die Möglichkeit, das Verhältnis von Bildung zu weiteren Bereichen einer Gesellschaft, beispielsweise zur Ökonomie, genauer zu bestimmen (Heinrich & Kohlstock, 2016).

Die soziale Ordnung im Bildungsbereich wird durch spezifische „strukturierte und strukturbildende Handlungen" (Altrichter, 2015, S. 38) (re)produziert. Handlungen und Struktur stehen in Relation und bedingen sich wechselseitig:

> [E]ine Akteurkonstellation [beinhaltet; E.M.] eine Struktur, die das Handeln der Akteure beeinflusst und durch das Handeln der Akteure wiederum verändert wird. Akteurkonstellationen stellen insofern Muster der sozialen Ordnungsbildung dar (Kussau & Brüsemeister, 2007, S. 27).

Dies erinnert an Bourdieus Theorie vom Habitus als strukturierende und strukturierte Struktur (Bourdieu, 1982, S. 279), bei der Handlungen und Struktur ebenfalls interdependent verbunden sind. Die soziale Ordnung im Bildungsbereich wird durch einen je spezifischen Habitus bestimmter Akteure und Akteurskonstellationen hergestellt und weitergegeben. Dieser beeinflusst die Interaktion der Akteure bzw. Akteurskonstellationen mit ihrer schulischen Umwelt. Änderungen dieser Umwelt, beispielsweise aufgrund von Reformen, wirken sich zunächst auf der Strukturebene aus bevor sie (selektiv) an die jeweiligen Bedingungen adaptiert und in Handlungen verschiedener Akteure und Akteurskonstellationen übersetzt werden. Diese Übersetzungsleistungen verweisen erneut auf Prozesse der Rekontextualisierung (Fend, 2008b; Kapitel 3.1). Die Handlungen müssen ihrerseits wiederum koordiniert werden, wobei sich diese Prozesse in den Formen Beobachtungs-, Beeinflussungs- oder Verhandlungskonstellationen bzw. deren Mischformen in Abhängigkeit von Verfügungsrechten manifestieren und mit den Modellen Hierarchie, Markt, Gemeinschaft und Netzwerke analysieren lassen. Gesamthaft ergeben diese Formen veränderbare national- und kulturspezifische „Governance-Regimes" (Kussau & Brüsemeister, 2007, S. 37ff.). Die Handlungskoordinationen können sowohl im Ausgleich kontroverser Interessen als auch in Blockaden münden, die in divergierenden Interessen, Rivalitäten oder Auseinandersetzungen um Deutungshoheit, Einflussnahme und Positionen ihre Ursache haben (Brüsemeister, 2010, S. 14f.; van Ackeren, Brauckmann, & Klein, 2016, S. 31). Zudem stehen die verschiedenen Akteure in Interdependenz, die „in rechtlich normierte, organisatorische und kulturelle Bedingungen eingebettet" (Kussau & Brüsemeister, 2007, S. 28) ist. Ihre Ausgestaltung unterliegt dem Einfluss von Regeln, Normen und Ressourcen der Akteure (Bormann & Hamborg, 2015, S. 296; auch Kussau & Brüsemeister, 2007, S. 30ff.).

Ob und wie Reformen wirken, ob sie in die gewünschte Richtung weisen und ob sie intendierte Veränderungen hervorrufen,

> entscheidet sich erst durch die Benutzung dieser ´Struktur- und Handlungsangebote´ und durch die Art und Weise, wie die verschiedenen Akteure (z. B. Lehrer/innen, Schüler/innen, Schulleitung und -aufsicht, Eltern) ihre Handlungen angesichts der Innovation […] neu ausrichten und koordinieren (Altrichter, 2015, S. 38f.).

Altrichter und Maag Merki (2016a, S. 16) differieren die Wirkungen erstens nach Effekten erster und zweiter Ordnung. Effekte erster Ordnung umfassen Modifikationen der Strukturen und Handlungen. Als Effekte zweiter Ordnung gelten Transferwirkungen von Reformen auf andere Bereiche des Systems. Ein zweites Unterscheidungsmerkmal stellen generelle und spezifische Effekte dar, wobei mehrere spezifische Effekte in Kombination generelle Effekte entwickeln können. Drittens gehen Reformen bei verschiedenen Akteuren mit unterschiedlichen Effekten einher.

Damit steht die Frage nach (politischer) Intentionalität und Transintentionalität beziehungsweise nach nicht-intendierten Auswirkungen intendierter Handlungen im Fokus von Educational Governance Analysen (Altrichter, 2015, S. 38ff., 56; Baum, 2014, S. 248; Kussau & Brüsemeister, 2007, S. 43). Vorgaben und Reformen sollen nicht nur zu oberflächlichen Veränderungen von Handlungen ganzer Akteursgruppen, sondern auch zu tiefergreifenden, individuell unterschiedlichen Änderungen von Wahrnehmungen, Denkschemata und Einstellungen führen (Koch, 2009, S. 134f.; auch Baum, 2014, S. 248).

> Aus steuerungstheoretischer Perspektive stellen die Einstellungen von Lehrkräften das Nadelöhr dar, durch das die *Steuerungspraxis* in und von Schule gelangen muss, um in *Handlungspraxis* transformiert zu werden (Koch, 2009, S. 135; Hervorhebungen im Original).

Neue Strukturen und neue Akteure treffen auf alte Strukturen und alte Akteure, die ihnen zum Teil nicht das Feld überlassen und etwas verändern, sondern Bestehendes bewahren wollen (Bennewitz, 2008, S. 257; Brüsemeister, 2010, S. 14; Kussau & Brüsemeister, 2007, S. 42f.; auch Helbig & Nikolai, 2015, S. 300).

Einige Antworten auf die Frage intendierter und nicht-intendierter Auswirkungen der Implementation zentraler Abiturprüfungen liefert die bisherige Forschung (Kapitel 4). Darüber hinaus untersucht die vorliegende Arbeit längerfristige Effekte der Modifikation des Prüfsystems. Zwar stellt das Handeln der Lehrpersonen nicht direkt den Gegenstand der Analysen dar, jedoch liefern die Befunde Hinweise auf potentielle Handlungen und Koordinationsprozesse.

3.4 Reformen und Innovationen im Bildungssystem

Bei Reform, Innovation, Implementation und Schulentwicklung handelt es sich nicht um trennscharfe Begriffe. Vielmehr verweist die Verwendung auf verschiedene Phasen und Akzentuierungen der Gestaltung und Steuerung des Schulsystems (Rürup, 2011; auch Dalin, 1999).

Reform

Aus Perspektive der Erziehungswissenschaft rekurriert der Begriff Schulreform einerseits auf die Reformpädagogik, andererseits auf Gestaltungs- und Veränderungsbemühungen des Schulsystems seitens des Staates und der Politik – mit Rückgriffen auf die Wissenschaft zur Legitimation – sowie auf das Scheitern der Politik, unter anderem aufgrund von Problemen bei der Implementation (Rürup, 2011, S. 17; Rustemeyer, 2009, S. 49; auch Emmerich & Maag Merki, S. 3). Veränderungen der Strukturen versteht Holtappels (2014, S. 20f.) als äussere Reform, die zu Veränderungen der Schulkultur als einer inneren, „echten" Reform führen können. Helbig und Nikolai definieren *Reformen* akteursunspezifisch als Reaktionen „auf ein von den Akteuren wahrgenommenes Problem" (Helbig & Nikolai, 2015, S. 138) – auf den Bildungsbereich übertragen als Problem eines mit Bildung befassten Akteurs. In eine ähnliche Richtung weist die Feststellung von Rürup über *Bildungsreform* als

> nicht mehr vorab und politisch-administrativ durchkonzipiert und als nur noch zu implementierendes Programm […], sondern ‚lediglich' als Konzept von Zielsetzungen und Schwerpunkten der Schulsystementwicklung […], die regional und einzelschulisch konkretisiert und realisiert werden sollen (Rürup, 2011, S. 20).

Dadurch, dass Reformen nicht mehr ausschliesslich „von außen an die Schule herangetragen" (Rürup, 2011, S. 17) werden, zeichnet sich auch ein Wandel und eine Öffnung der Akteure oder Akteursgruppen, die Veränderungen initiieren, ab.

Für einen Überblick über Reformen bezüglich Strukturen, Inhalte und deren Kontrolle im Bildungsbereich in den deutschen Bundesländern von den 1950er Jahren bis zu den 2000er Jahren sei auf das Werk von Helbig und Nikolai (2015) verwiesen. Dort werden unter anderem auch die Veränderungen bei den Abiturprüfungen nachgezeichnet, differenziert nach verschiedenen Typen von Bildungssystemen. Das im Fokus dieser Arbeit stehende Bremen wird dem modernisiert-destandardisierten Typ zugeordnet mit „Tendenz in Richtung einer stärkeren Standardisierung" (Helbig & Nikolai, 2015, S. 289). Insgesamt bescheinigen die Autoren Bremen relative „Reformfreudigkeit", wobei sich Phasen mit mehreren Reformen mit Phasen geringer Reformtätigkeiten abwechseln (Helbig & Nikolai, 2015, S. 142). Ein Grund dafür dürfte unter anderem sein, dass Reformtätigkeiten zwar nicht ausschliesslich, jedoch in hohem Masse, von politischen Akteuren initiiert werden. Die politischen Akteure bzw. Akteurskonstellationen ändern sich jedoch aufgrund von Wahlen. „Neue politische Konstellationen und programmatische Neuorientierungen bringen neue Politiken. […] Eine unvollendete Reform folgt der nächsten. Die Absichten und Richtungen sind oft widersprüchlich, jedenfalls unstet." (Meyer-Hesemann, 2010, S. 87) Dies steht der eingangs zitierten Erkenntnis, dass Reformen viel Zeit benötigen, entgegen. Hinzukommen Änderungen in den Einstellungen der politischen Akteure aufgrund von unvorhergesehenen Ereignissen.[6]

> Es ist eine für Historiker und Sozialwissenschaftler altbekannte Tatsache, dass einschneidende und rasche gesellschaftliche Veränderungsprozesse in ihrem Verlauf nicht vollständig antizipierbar sind und stets zu unerwarteten Entwicklungen – sowohl zu unerwarteten Entdeckungen von Erneuerungspotentialen als auch zu unerwarteten Schwierigkeiten auf dem Wege der Erneuerung – führen (Schütze & Breidenstein, 2008, S. 10).

[6] Als Beispiel sei auf die Änderung der Haltung der CDU in Deutschland zur Kernenergie nach den Vorfällen in Fukushima und der damit einhergehenden kritischen Sichtweise von Atomenergie vieler Bürgerinnen und Bürger verwiesen. Der Bildungsbereich mag für derartige „180°-Wendungen" weniger anfällig erscheinen, doch auch hier finden sich Beispiele, etwa die Einführung der gymnasialen Schulzeitverkürzung von neun auf acht Jahre, die mittlerweile in einigen Bundesländern wieder teilweise oder ganz zurückgenommen wurde (Ivanov, Nikolova, & Vieluf, 2016; Kühn, 2016).

Hingegen zeichnet Reynolds (2014), zumindest für die USA, ein kritisches Bild von Reformen. Er konstatiert trotz aller Diskussionen und politischer Eingriffe lediglich eine Art „Recycling" wiederkehrender Ideen oder Ideologien statt grundlegender Reformen im Bildungsbereich.

Innovation

Das von Reynolds (2014) angedeutete Spannungsfeld von „wahren" Neuerungen bzw. Neuheit und „neuem Wein in alten Schläuchen", das heisst Neuerung bzw. Erneuerung, bestimmt auch den Begriff *Innovation*. Dieser wurde zunächst in den Wirtschaftswissenschaften verwendet, später von der Politik sowie anderen Wissenschaftsdisziplinen, auch der Erziehungswissenschaft und Bildungsforschung, aufgegriffen und stellt mittlerweile ein interdisziplinäres Forschungsfeld dar. Innovation beinhaltet eine normative Komponente, die eine Verbesserung des aktuellen Zustandes impliziert (Rürup, 2011, S. 10ff.; auch Bormann, 2011, S. 54; Rogers, 2003, S. 15; van Ackeren et al., 2011, S. 176).

- Innovationen können grundlegende Neuerungen sein oder Verbesserungen von Verfahren oder Strukturen,
- Innovationen sind neu im radikalen Sinne des ‚zum ersten Mal in der Welt sein', oder sie sind neu für das System, das diese Innovation einführt, und
- Innovationen sind soziale Prozesse (Blättel-Mink & Menez, 2015, S. 34).

Innovationen bezeichnen unterschiedliche Zustände und Prozesse, wobei es Rogers (2003, S. 12) zufolge entscheidend ist, dass ein Akteur oder eine Akteursgruppe eine Idee für neu hält und nicht, ob sie tatsächlich neu ist. Innovationen lassen sich unterscheiden nach ihrem Grad an wahrgenommener Überlegenheit gegenüber vorherigen Ideen, an Übereinstimmung mit den Werten, Erfahrungen und Bedürfnissen betroffener Akteure, an Komplexität, an Möglichkeiten, sie in kleinem Rahmen auszuprobieren und an Sichtbarkeit ihrer Ergebnisse bzw. Folgen (Rogers, 2003, S. 15f.).

Je nach Tradition werden Innovationen als lineare (top-down-)Prozesse, zirkuläre Prozesse mit Fokus auf den Voraussetzungen der Adressaten, den Verlauf und offenem Ausgang oder als soziale Praxis mit reflexiven Prozessen verstanden (Blättel-Mink & Menez, 2015, S. 61; Bormann, 2011, S. 55ff.). Der Blick auf die Adressaten und deren Bedingungen verweist erneut auf Rekontextualisierungsprozesse (Kapitel 3.1) sowie die im Fokus der Educational Governance stehenden Handlungskoordinationen (Kapitel 3.3).

Zudem spielen Entscheidungsspielräume und Entscheidungsbefugnisse über die Annahme und Einführung einer Innovation eine Rolle. Die Einführung des Zentralabiturs in Bremen gehört demnach in die Kategorie „authority innovation-decisions": Eine kleine Gruppe von Menschen mit Macht, Status oder technischer Expertise fällt die Entscheidung für oder gegen eine Innovation, nicht jedoch die einzelnen Individuen innerhalb eines Systems (Rogers, 2003, S. 28f.).

Bormann zufolge stellen Innovationen, genauer der Transfer von Innovationen, „Wissenspassagen", das heisst „kollektive, wissensbasierte Phänomene mit zeitlich, räumlich, sachlich, sozial und kognitiv differenzierenden Implikationen" (Bormann, 2011, S. 53) dar. Sowohl individuelle oder kollektive Akteure als auch Systeme oder deren Interaktion können Innovationen initiieren. Innovationen sind demnach Prozesse, „in denen Wissen für das Verstehen einer diskursiv übermittelten Veränderungsabsicht generiert, aktiviert und angewendet wird und in deren Zuge neue Ordnungen von Wissen entstehen" (Bormann, 2011, S. 57). Das implizite wie explizite Wissen individueller und kollektiver Akteure ist jeweils kontextabhängig, weshalb „Innovationen als Gegenstände und gleichermaßen als Resultate von Wissensarbeit" (Bormann, 2011, S. 58f.) zu betrachten sind. Der Verweis auf die notwendige Analyse der „‚inneren' Verarbeitung eines von ‚außen' initiierten Vorgangs der Sinnstiftung" (Bormann, 2011, S. 59) erinnert erneut an die Definition von Sozialisation als die „produktive Verarbeitung der inneren und äußeren Realität" (Hurrelmann, 2006, S. 11). Die Bedeutung des Kontextes der Wissensgenerierung lässt sich vor diesem Hintergrund als eine Form der beruflichen individuellen wie kollektiven Sozialisation verstehen. Im Gegensatz zu Reformen beinhalten Innovationen erstens stärker die Möglichkeit von bottom-up-initiierten Veränderungen. Dies hängt auch damit zusammen, dass, zweitens, Reformen auf politische Steuerungsbemühungen verweisen, deren konzeptionelle Neuheiten als Innovation bezeichnet werden (Rürup, 2011, S. 18ff.). Dieser Unterscheidung folgend, wird die Neuausrichtung der Prüfungsorganisation von dezentralen zu *zentralen Abiturprüfungen* als von Seiten der Bildungspolitik (Makroebene) initiierte Reform verstanden. Auf Seiten der einzelnen Akteure und Akteursgruppen kann das Zentralabitur als Innovation erscheinen.

Akteure bzw. Akteursgruppen, welche in einen gesellschaftlichen Kontext eingebettet sind, initiieren sowohl Innovationen als auch Reformen. Sie entstehen im Rahmen der „Deutung und (Re-)Produktion"

(Blättel-Mink & Menez, 2015, S. 31) der gesellschaftlichen Bedingungen und Zusammenhänge. Sowohl Einstellungen der einzelnen Akteure als auch Merkmale des „Organisationsklimas bzw. der Organisationskultur sowie organisationsstrukturelle[r] Bedingungen" (van Ackeren et al., 2011, S. 179) sind von Bedeutung. Dies verweist zum einen erneut auf die Interdependenz von Strukturen und Handlungen (Kapitel 3.3) sowie auf die Mikro-, Meso- und Makroebene des Bildungssystems (Kapitel 3.1). Zum anderen zeigen die Forschungen von Buske (2014) zur kollektiven Innovationsbereitschaft von Lehrpersonen, dass über die einzelnen Akteure hinaus auch deren Konstellation bzw. Gruppierung eine Rolle spielt. *Folgen* von Innovationen wie Reformen können auf jeder Ebene differenzielle erwünschte und unerwünschte, direkte und indirekte, erwartete und unerwartete, generelle und spezifische *Auswirkungen* sein (Altrichter, 2015, S. 39f.; Altrichter & Maag Merki, 2016a, S. 16; Kussau & Brüsemeister, 2007, S. 43; Rogers, 2003, S. 30f.). Die Spannweite der Folgen lässt sich unter anderem auf die relativ grossen Handlungsfreiheiten der einzelnen Schulen und Akteure bei gleichzeitig geringen Anreiz-, Kontroll- und Sanktionsmöglichkeiten seitens der Bildungspolitik zurückführen (Baum, 2014, S. 248; Baumert, 2016, S. 222; van Ackeren et al., 2015, S. 105f.; van Ackeren et al., 2011, S. 174).

Implementation

Reformen und Innovationen müssen *implementiert*, das heisst in der Praxis eingeführt und umgesetzt werden (Berner, Oelkers, & Reusser, 2008, S. 224). Dafür müssen vor allem Lehrpersonen Zeit und Gelegenheiten für kollegiale Diskussionen haben. Die Passung einer Reform oder Innovation zur Schulkultur der Einzelschule ist dabei für die Entwicklung der Schule entscheidend (Hopkins, Ainscow, & West, 1994; auch Dalin, 1999). Reynolds geht in eine ähnliche Richtung und erachtet die Implementierung als nicht ausreichend, sondern fordert die Institutionalisierung der Veränderungen: „Change is only successful when it has become part of the natural behaviour of teachers in the school. Implementation by itself is not enough." (Reynolds, 2005, S. 20; Rogers, 2003: „routinization") Die verschiedenen Ebenen, zahlreichen Akteure und Akteurskonstellationen, Rekontextualisierungsprozesse und Handlungskoordinationen (Kapitel 3.1, Kapitel 3.3) sowie die „Kommunikationswege zwischen den Initiatoren einer Reform und denjenigen, die sie in der Praxis umsetzen sollen" (Zeitler, 2012, S. 23) stellen eine Herausforderung für diesen Prozess dar. Widersprüche und Konflikte aufgrund von Werte- und Macht-Barrieren, Problemen bei der Interpretation und Realisierung oder Eigenschaften der Akteure scheinen Bestandteil von

Reformen bzw. Innovationen zu sein (Holtappels, 2014, S. 21; auch Bennewitz, 2008, S. 248; Dalin, 1999; Spillane, Reiser, & Reimer, 2002). Van Ackeren et al. sprechen von

> 'Implementationsbrüchen', die sowohl zwischen den verschiedenen strukturellen Ebenen im Schulwesen als auch innerhalb einer Ebene möglich und erwartbar sind, so dass die zeitliche und strukturelle Stabilität und Kontinuität der Implementierungsprozesse stets gefährdet ist (van Ackeren et al., 2011, S. 174).

Dies führt letztlich dazu, dass sich der Implementationsprozess sowohl auf unterschiedlichen Ebenen als auch zwischen verschiedenen Akteuren und Akteurskonstellationen einer Ebene unterschiedlich entwickelt, was eine zusätzliche Dimension der Komplexität und Differenz ins System einführt. „Das Gesamtbild der Erscheinungen zwischen dem Pol des Beharrens und dem des Veränderns ist also äußerst komplex und oftmals auf den ersten Blick nicht transparent" (Schütze & Breidenstein, 2008, S. 11).

Insbesondere Lehrpersonen sind von Reformen betroffen. Sie sollen nicht nur neue Vorgaben umsetzen, teilweise unter neuen und ungewohnten Rahmenbedingungen, das heisst ohne Handlungsroutinen, sondern auch durch ihr Handeln Verbesserungen in den Lernergebnissen der Schülerinnen und Schüler hervorrufen. Letztere bilden schliesslich die Grundlage für die Bewertung von Erfolg oder Misserfolg der Reform. Reformen oder Teile von Reformen können von Lehrpersonen als positive, sinnstiftende Herausforderung und Entlastung oder als negative Anstrengung und Belastung erlebt werden (Bennewitz, 2008, S. 247f.; Spillane et al., 2002; van Veen, Sleegers, & van de Ven, 2005). Dies hängt nicht nur von individuellen und kollektiven Einstellungen, Erfahrungen und Deutungsmustern ab, sondern auch von Bedingungen der Einzelschule und der gesellschaftlichen Einbettung der Reform, was erneut auf die verschiedenen Ebenen des Bildungssystems und deren Interdependenzen verweist (z. B. Altrichter, 2015; Fend, 2008b; Hamilton et al., 2008; Herman, 2005; Schmidt & Datnow, 2005). Zur Beschreibung von *sense-making*-Prozessen während der Implementation berücksichtigen Spillane et al. (2002) die kognitiven Prozesse, die zum Verständnis der eigenen Handlungen beitragen und zur Veränderung von Einstellungen und Überzeugungen führen können. Unterschieden werden erstens individuelle kognitive Prozesse (individuelles sense-making) der Reaktion auf und Interpretation von

Reforminitiativen, welche von Vorwissen, Vorerfahrungen, Einstellungen, Erwartungen und Emotionen abhängen. Zweitens ist der Kontext mit seinen jeweiligen Normen, Regeln, Strukturen und Möglichkeiten der Interaktion verschiedener Akteure und Akteursgruppen entscheidend (soziales sense-making). Demnach enthält der Kontext eine formale, informelle sowie historische Dimension und zwar jeweils auf der Makro-, Meso- und Mikroebene, welche wiederum in Beziehung stehen. Drittens wirken Vorstellungen und Ideen (externe Repräsentationen) über Veränderung und Wandel, die seitens der Politik oder Verwaltung kommuniziert werden. Das Modell kognitiver Prozesse ist nicht-linear. Es verdeutlicht die Komplexität der Einflüsse auf den Prozess der Implementation von Reformen oder Innovationen und zeigt, wie voraussetzungsreich dieser ist. Folgen können vielfältige absichtliche und unabsichtliche Fehlinterpretationen der Intentionen und Ideen sein, welche die Implementation verhindern. Insbesondere, wenn sie den Unterricht als Kern des Lehrberufs betreffen. (Spillane et al., 2002, S. 388ff.)

> In sum, our usual approach to processing new knowledge is a conserving process, preserving existing frames rather than radically transforming them. New ideas either are understood as familiar ones, without sufficient attention to aspects that diverge from the familiar, or are integrated without restructuring of existing knowledge and beliefs, resulting in piecemeal changes in existing practice. (Spillane et al., 2002, S. 398)

Die Berücksichtigung individueller kognitiver Prozesse und Dispositionen für die längerfristige und intendierte Anwendung von Innovationen findet sich auch im *Pädagogischen Akzeptanzmodell* von Ziegelbauer (2015, S. 153ff.), das in drei Phasen die Entstehung von Akzeptanz einer Innovation in den Blick nimmt. In der präaktionalen Phase steht die Einstellungsakzeptanz, in der aktionalen Phase die Verhaltensakzeptanz und in der postaktionalen Phase die Nutzungsakzeptanz im Vordergrund, die sich wiederum auf die Einstellungsakzeptanz auswirkt. Für die Implementation bestehen folgende Kriterien: „Beschaffenheit der Innovation, Rahmenbedingungen des Systems und Voraussetzungen bei den Anwendern" (Ziegelbauer, 2015, S. 155), wobei das Modell letzteren besonderes Gewicht verleiht.

Die bei Spillane et al. (2002) angesprochene besondere Bedeutung von Interaktionen findet sich ebenfalls bei Penuel, Frank, Sun und Kim (2012, S. 184f.), die in diesem Zusammenhang vom *sozialen Kapital von Lehrpersonen* sprechen. Das soziale Kapital umfasst Ressourcen und Wissen, auf das Lehrpersonen

aufgrund ihrer durch Interaktionen entstandenen netzwerkartigen Verbindungen mit (Subgruppen von) anderen Lehrpersonen zurückgreifen können. Diese Verbindungen können einerseits einen normativen und sozialen Druck erzeugen, Reformen in der Schule zu implementieren und Handlungen entsprechend anzupassen. Dies ist vorrangig der Fall, wenn die Lehrpersonen häufig interagieren. Andererseits liefern die Verbindungen Zugang zu Wissen, das für die Implementation von Reformen nötig ist.

Die Implementation des Zentralabiturs bedeutet dementsprechend nicht lediglich eine blosse Umstellung der Prüfungsorganisation am Ende der Sekundarstufe II, während alles andere unverändert bleibt, sondern Veränderungen von Strukturen, Handlungen und Interaktionen verschiedener Akteure und Akteursgruppen, vor allem bei den Lehrpersonen, Schülerinnen und Schülern.

3.5 Schulentwicklung und Schuleffektivität

Eng verbunden mit der (zentralen) Steuerung von Bildungs- bzw. Schulsystemen und damit einhergehenden Reformen und Innovationen sind Fragen der (dezentralen) Entwicklung[7] hin zu einer „guten Schule" und der Effektivität von Schule (Berkemeyer, Bos, & Gröhlich, 2010, S. 147).

Schulentwicklung

Bei *Schulentwicklung* sind zwei Perspektiven zu unterscheiden: einerseits die Planung und Bereitstellung eines staatlichen Schulangebots für alle Schülerinnen und Schüler (Makroebene) und andererseits Veränderungen im Sinne eines „systematischen, intentionalen Qualitätsentwicklungs-Prozesses" (van Ackeren et al., 2015, S. 189) innerhalb von Einzelschulen (Mesoebene), die schulische wie wissenschaftliche Akteure durch Entwicklung, Umsetzung und Überprüfung gemeinsamer Ziele generieren (Bryk, 2010; Dalin, 1999; Hallinger & Heck, 2011; Hopkins et al., 1994; Rürup, 2011, S. 15f.; van Ackeren et al., 2015, S. 189; auch Rolff, 2010, S. 36). Die zwei Perspektiven lassen sich jedoch nicht trennscharf abgrenzen, zumal die Makroebene weiter in eine normative und eine organisatorische Komponente unterteilen werden kann (Emmerich & Maag Merki, 2014, S. 4). Vielmehr bildet die Makroebene die Umwelt für die Mesoebene (Maag Merki & Werner, 2013, S. 296) – und im Sinne des Mehrebenencharakters auch für die Mikroebene.

[7] Emmerich und Maag Merki (2014) verweisen auf die nötige, jedoch nicht stets offensichtliche Unterscheidung zwischen Entwicklung als geplanter Veränderung und Evolution als willkürlicher Veränderung.

> Angesichts der nicht aufhebbaren Differenz von zentralstaatlicher Schulpolitik und regionaler Schulentwicklung schufen die Regulierungen der Bildungsadministration gleichsam durch regulierte Deregulierung ein strukturelles Bedingungsgefüge, innerhalb dessen sich realiter unterschiedlichste regionale Schulentwicklungs- und Versorgungsmuster herausbilden konnten (Herrmann, 2009, S. 65).

Rolff konzipiert Schulentwicklung als Kombination aus Organisations-, Unterrichts- und Personalentwicklung, welche untereinander verbunden sind. Er differenziert drei Ebenen: Schulentwicklung erster Ordnung umfasst die „bewusste und systematische Weiterentwicklung von Einzelschulen" (Rolff, 2010, S. 36), welche mittels Schulentwicklung zweiter Ordnung bzw. institutioneller Schulentwicklung zu Lernenden Schulen werden sollen. Schulentwicklung dritter Ordnung oder komplexe Schulentwicklung beinhaltet die Steuerung des Gesamtzusammenhangs (Rolff, 2010, S. 36). Die Schulentwicklung der ersten und zweiten Ordnung lässt sich der Mesoebene, die der dritten Ordnung der Makroebene zuordnen.

> Gründe für diese Veränderungsprozesse wie auch für die Notwendigkeit systematischer Veränderungen können Defizite der Schule selbst, aber auch Veränderungen im schulischen Umfeld oder neue Vorgaben der Bildungspolitik und Bildungsverwaltung sein (Bischof, 2017, S. 30).

Über die „Richtung" von Entwicklungsanregungen besteht Uneinigkeit: Einerseits wird davon ausgegangen, dass Entwicklungen extern wie intern angeregt werden können (Baum, 2014, S. 248; Emmerich & Maag Merki, 2014, S. 2f.; Fullan, 1999; Thoonen, Sleegers, Oort, & Peetsma, 2012, S. 442ff.). Den Lehrpersonen kommt bei der Umsetzung der Entwicklungen eine entscheidende Rolle zu (Baum, 2014, S. 248; Sahner, 2008): „Lehrer sind Vermittler des Bildungswandels und des gesellschaftlichen Fortschritts" (Fullan, 1999, S. 31). Dagegen verortet Bellmann (2016, S. 26ff.) interne Entwicklungsanreize bei der Schulentwicklung und externe Entwicklungsanreize bei der Neuen Steuerung (Kapitel 3.2). Maag Merki und Werner (2013) sehen indes darin einen neuen Typus von *Schulentwicklungsforschung*, der sich mit den indirekten Wirkungen externer Steuerungsinstrumente auf die Lernergebnisse der Schülerinnen und Schüler befasst. Die Vermittlung erfolgt „über die Weiterentwicklung der

schulischen und unterrichtlichen Prozesse sowie der Professionalisierung der Lehrpersonen, d.h. über Schulentwicklungsprozesse" (Maag Merki & Werner, 2013, S. 301). Vor dem Hintergrund der *standards-based accountability* in den USA unterscheidet Schraw (2010, S. 73) einerseits Schulentwicklung bzw. *school improvement* als negativen Konsequenzen (z. B. Sanktionen) vorbeugende Massnahmen und andererseits als Folge von Sanktionen. Der Output bildet demnach den entscheidenden externen Anreiz für Veränderungen, die den Unterricht und in dessen Folge die Testergebnisse und die damit einhergehenden „school`s chances of a favorable accountability review" (Schraw, 2010, S. 73; auch Forte, 2010) verbessern sollen. Da interne Entwicklungsanlässe nicht berücksichtigt werden, liegt ein engeres Verständnis von Schulentwicklung vor.

Im Fokus der *Schulentwicklungsforschung* wie auch ihrer theoretischen Fundierung stehen sowohl (soziale) Strukturen als auch (kollektive) Handlungsprozesse und Wirkungen von Schulentwicklung, oft unter Einbezug mehrerer Zeitpunkte (Emmerich & Maag Merki, 2014, S. 19f.; Feldhoff, Bischof, Emmerich, & Radisch, 2015, S. 65ff.). Es geht demnach um die „Identifizierung von Faktoren, die systematische, zielgerichtete und nachhaltige Veränderungsprozesse der Schule ermöglichen, sowie [die; E.M.] Beschreibung dieser Prozesse" (Bischof, 2017, S. 33). Holtappels unterscheidet drei Arten von Schulentwicklungsforschung:

1. Analyse intendierter und systematischer Schulentwicklungsstrategien
2. Analyse des Zeitverlaufs von Entwicklungen ohne intendierte und systematische Schulentwicklungsstrategien
3. Analyse des Zeitverlaufs von Reformen, Innovationen oder Schulentwicklungs-programmen und -strategien (Holtappels, 2014, S. 12).

Die Analyse der Implementation zentraler Abiturprüfungen über einen Zeitraum von fünf Jahren kann als eine Mischform der zweiten und dritten Kategorie gesehen werden. Sie stellt eine Reform der Prüfungsorganisation dar und ist zugleich mit Unterrichtsentwicklung verbunden (Kapitel 2.2). Da jedoch fraglich ist, wie systematisch und „entwicklungsstrategisch" die Intentionen seitens der Bildungspolitik waren, sind die Analysen eher der zweiten als der dritten Kategorie zuzuordnen.

Der Konzeption von Schulentwicklung als Verschränkung von Unterrichts-, Organisations- und Personalentwicklung (Rolff, 2010) und dem Verständnis von Schulentwicklungsforschung von Maag Merki und Werner (2013) folgend, kann das mit der bildungspolitisch initiierten Reform der Prüfungsorganisation von dezentralen zu zentralen Abiturprüfungen eingeführte externe Steuerungsinstrument auf Ebene der Einzelschule Schulentwicklungsprozesse auslösen (Bischof, 2017). In dem Fall stehen bisherige gemeinsame Ziele, Strukturen und Prozesse auf dem Prüfstand, müssen ausgehandelt, koordiniert und gegebenenfalls neu entwickelt werden, um die Schul- und Unterrichtsqualität unter den veränderten Rahmenbedingungen aufrecht zu halten oder zu verbessern. Beispielsweise kann die Vorgabe externer Abituraufgaben und Korrekturkriterien zu einer Diskussion darüber führen, welche pädagogischen Ziele und Leitideen die einzelnen Lehrpersonen ebenso wie die gesamte Schule mit der Verwendung bestimmter Aufgabenformate und Bezugsnormen bei der Benotung verbinden. Ebenfalls möglich ist etwa die Kooperation von Lehrpersonen mit dem gleichen Kurs (z. B. Leistungskurs Mathematik) bei der Planung von Unterricht hinsichtlich der extern vorgegebenen inhaltlichen Schwerpunktthemen sowie bei Prüfungen und deren Korrektur.

Das Potenzial des Zentralabiturs, Schul- und Unterrichtsentwicklung anzustossen, sehen Kahnert, Eickelmann, Lorenz und Bos allerdings eher eingeschränkt, da „die Ergebnisse des Zentralabiturs in den meisten Bundesländern mit wenig unterrichtsbezogenem Informationsgehalt, z. B. in Form von Rückmeldungen von Notenmittelwerten auf Schulebene, veröffentlicht" (Kahnert et al., 2015, S. 89) werden und nicht in detaillierter Form wie bei anderen Instrumenten der outputorientierten Steuerung, beispielsweise bei Vergleichsarbeiten und Lernstandserhebungen (van Ackeren & Bellenberg, 2004). Geht man jedoch nicht von den Ergebnissen der Schülerinnen und Schüler im Abitur aus, sondern von den Prüfungen an sich, rückt der ihnen vorgelagerte Unterricht, der auf die allen Beteiligten unbekannten Prüfungsthemen vorbereiten soll, in den Fokus. Die von Kahnert et al. aufgeworfene Frage, inwiefern durch die „Vorgaben für Abiturprüfungen der vorgelagerte Unterricht beeinflusst wird" (Kahnert et al., 2015, S. 90) bzw. beeinflusst werden kann, beantworten die theoretischen Ausführungen zu teaching to the test und test preparation (Kapitel 2), die Ableitung des eigenen Modells (Kapitel 3.6) sowie empirische Befunde (Kapitel 4).

Schule als Organisation

Mit Schulentwicklung ist die Diskussion verbunden, ob Schulen Organisationen oder Lernende Organisationen sind (Argyris & Schön, 1999; Dalin, 1999; Dubs, 2010; Feldhoff, Gromala, & Brüsemeister, 2014: „organisationales Lernen"; Fullan, 1999: „Lernende Unternehmen"; Holtappels, 2010a; Hopkins et al., 1994; Rustemeyer, 2009; Schratz & Steiner-Löffler, 1999). Das Schulsystem ist gemäss Rolff (1993, S. 121) „die zahlenmäßig größte, technisch einfachste und sozial komplizierteste Organisation mit dem qualifiziertesten Personal". Rolff arbeitet sechs Merkmale heraus, die die Schule von anderen Organisationen unterscheidet, weshalb die Schule eine „besondere soziale Organisation" (Rolff, 1993, S. 121) ist:

1. Bildungsauftrag – Vermittlung von Inhalten
2. pädagogischer Bezug – begrenzte Technologisierbarkeit
3. Schüler stehen im Mittelpunkt – Fallverstehen als Grundlage pädagogischen Handelns
4. Lehrer als unvollendete Professionelle – gebrochene Kontrolle
5. Arbeitsteilung – gefügeartige Kooperation
6. Erziehung zur Selbsterziehung – Reflexivität der Ziele (Rolff, 1993, S. 125ff.).

Diese Charakteristika von Schulen lassen vor allem Schwierigkeiten, Hürden und Grenzen der direkten Steuerung und Steuerbarkeit ersichtlich werden, wie bereits in den vorherigen Kapiteln ausgeführt. Im Gegensatz zu Rolff (1993) und etlichen weiteren Autoren wie etwa Bryk (2010: „complex organizations"), Emmerich und Maag Merki (2014), Muders (2016), Rustemeyer (2009: „pädagogische Organisationen") oder van Ackeren et al. (2015) kommt Böttcher (2012, S. 39f.) zu dem Schluss, dass Schulen keine Organisationen darstellen, da sie wesentliche Merkmale sozialer Organisationen nicht erfüllen.

Unabhängig davon, ob Schulen als Organisationen verstanden werden, handelt es sich bei Schulentwicklung um komplexe soziale Prozesse:

> the complexity of school improvement as a social process can be described by six characteristics: the longitudinal nature; the indirect nature; the multilevel phenomenon; the reciprocal nature; differential development and nonlinear effects; and the variety of meaningful factors (Feldhoff, Radisch, & Bischof, 2016, S. 213).

Schuleffektivität

Die Frage, wie und unter welchen Bedingungen sich Schulen entwickeln, verweist auf die Frage des „Ziels" der Entwicklung: Was macht eine „gute" Schule aus und wie können mehr Schulen zu „guten" Schulen werden (Reynolds et al., 2014, S. 197; Scheerens, 1992, S. 11f.) und damit auf das theoretische und empirische Feld der *Schuleffektivität*, auch Schulwirksamkeit oder educational effectiveness (inklusive school effectiveness und teacher effectiveness). Der Fokus liegt auf den Faktoren, welche die kognitiven wie nicht-kognitiven Lernergebnisse der Schülerinnen und Schüler, abgesehen von deren individuellen Dispositionen, auf verschiedenen Ebenen beeinflussen. Schulische Faktoren können direkte und indirekte Effekte erzielen, beispielsweise vermittelt über den Unterricht oder die Interaktion zwischen Lehrperson, Schülerinnen und Schülern. Diese Effekte können kausal, linear oder nicht-linear sein. (Muijs, Campbell, & Kyriakides, 2005; Muijs et al., 2014; Reynolds, 2005, S. 13; Reynolds, Teddlie, Chapman, & Stringfield, 2015; Scheerens, 1992; Teddlie & Reynolds, 2000)

Gemäss des *context-input-process-output-outcome model of schooling* (Scheerens, 1990, S. 63) beeinflussen Indikatoren der Kontextebene (Anforderungen, Rahmenbedingungen der Schule, Bildungsdaten auf der Makroebene) das Zusammenspiel von Input (Ressourcen, Ausbildung der Lehrpersonen), Prozessen (Curriculum, Organisation und Klima der Schule), Output (Leistungen) und Outcome (Beschäftigung, Verdienst). Im *integrated model of school effectiveness* differenziert Scheerens (1990, S. 73) die Indikatoren und Wirkungsrichtungen von Kontext, Input, Prozess und Output basierend auf Befunden der Schuleffektivitätsforschung aus und stellt die top-down-Beziehungen zwischen der Makro-, Meso- und Mikroebene deutlicher heraus. Prozessindikatoren kommt eine besondere Bedeutung für die Erklärung von Unterschieden im Output zwischen Schulen oder Schulsystemen zu. Darüber hinaus bilden sie Ansatzpunkte für Veränderungen mit dem Ziel, den Output zu verbessern (Scheerens, 1990, S. 69f.).

Die Mehrebenenstruktur (Kapitel 3.1) und der Einfluss verschiedener Indikatoren auf den Outcome von Schülerinnen und Schülern findet sich ebenfalls beim *dynamic model of educational effectiveness* von Creemers und Kyriakides (2008)[8], eine Erweiterung des *comprehensive model of school effectiveness* von Creemers (1994; n.d.). Grundannahme ist, dass sich Faktoren von vier Ebenen auf den kognitiven, affektiven und psychomotorischen Outcome sowie auf das von den Autoren nicht näher definierte „new learning" der Schülerinnen und Schüler auswirken.

[8] Deutsche Übersetzung in Feldhoff et al. (2015, S. 76)

- Kontext der Schule: nationale bzw. regionale Bildungspolitik, Evaluationen, Wertvorstellungen
- Schule: Politik der Einzelschule bezüglich des Unterrichts (z. B. Organisation, Hausaufgaben) und des Aufbaus einer „school learning environment" (z. B. Zusammenarbeit und Interaktion der Lehrpersonen, Wertvorstellungen), Handlungen für deren Verbesserung, Evaluation der Politik, Handlungen und school learning environment
- Klasse: Unterrichtsqualität (z. B. Strukturierung, Leistungsbeurteilung, Zeitmanagement)
- Schülerin bzw. Schüler (Verbindung einer soziologischen und psychologischen Perspektive): 1. zeitinvariante Variablen wie sozioökonomischer und soziokultureller Hintergrund, Geschlecht, Persönlichkeitsmerkmale; 2. veränderbare Variablen wie Lerngelegenheiten, Erwartungen, Motivation (Creemers & Kyriakides, 2008, S. 75ff.; 2010, S. 264ff.; Creemers et al., 2013b, S. 11ff.; Kyriakides, 2008, S. 442f.; Kyriakides, Creemers, Antoniou, & Demetriou, 2010, S. 808ff.).

Die Merkmale der Schülerinnen und Schüler stehen einerseits mit deren Outcome und andererseits mit der Ebene der Klasse in einem wechselseitigen Zusammenhang, das heisst sie wirken sich sowohl direkt als auch indirekt über den Unterricht aus. Letzteres gilt ebenfalls für Faktoren der Schulebene und Kontextebene (indirekt über Klassenebene und „doppelt" indirekt über Schulebene und anschliessend Klassenebene), wobei hier die Wirkungsrichtungen unidirektional top-down sind, da die höhere(n) Ebene(n) Umwelt der niedrigeren ist bzw. sind. Dass die Effekte der Kontext- und Schulebene in Abhängigkeit der aktuellen Situation der Einzelschule variieren können, ist ein zentraler Punkt des Modells. Ein weiterer ist die entscheidende Bedeutung, die der Klassenebene, das heisst den Lehr-Lern-Prozessen, trotz der Mehrebenenstruktur (Kapitel 3.1) zukommt. Zum Modell gehört die Annahme, dass sich jeder Faktor der Klassen-, Schul- und Kontextebene mittels der Dimensionen Häufigkeit, Fokus, Phase, Qualität und Differenzierung messen lässt. Dadurch wird insbesondere konkretisiert, wie Unterrichtsqualität verstanden wird. (Creemers & Kyriakides, 2008, S. 75ff.; 2010, S. 264ff.; Creemers et al., 2013b, S. 11ff.; Kyriakides, 2008, S. 442f.; Kyriakides et al., 2010, S. 808ff.)

Kritisch zu sehen ist, dass der individuelle Hintergrund und Persönlichkeitsmerkmale der Schülerinnen und Schüler ausführlich berücksichtigt werden, Merkmale der Lehrpersonen, die mit ihren Handlungen und ihrem Unterricht in Verbindung stehen, wie beispielsweise Einstellungen, Motivation oder Alter hingegen nicht.

> However, these characteristics should not be included in the models of educational effectiveness. This argument is not only supported by the fact that these teacher characteristics were not found to be related to achievement. It is also argued here that the models of effectiveness should concentrate on the teaching activities teachers perform in order to initiate, promote, and evaluate student learning. (Kyriakides, 2008, S. 442)

Viele der theoretisch formulierten Zusammenhänge und Wirkungen konnten in empirischen Studien bestätigt werden (z. B. Creemers et al., 2013b; Kyriakides, 2005, 2008; Kyriakides, Christoforou, & Charalambous, 2013).

Ziel der *Schuleffektivitätsforschung* (educational effectiveness research) ist es

> to investigate all the factors within schools in particular, and the educational system in general, that might affect the learning outcomes of students in both their academic and social development, which means it encompasses a wide range of factors such as teaching methods, the organization – formally and informally – of schools, the curriculum, the role of leadership, and the effects of educational 'learning environments' in general, whether schools, districts, or nations (Reynolds et al., 2014, S. 197).

Kyriakides zufolge geht es jedoch nicht allein um die Analyse der genannten Faktoren, sondern vorrangig darum, sie mittels Schulentwicklungsprojekten zu implementieren bzw. zu verändern (Kyriakides, 2008, S. 441). Hier deutet sich der Versuch einer Verbindung von Schulentwicklung und Schuleffektivität an.

Überträgt man die Modelle der Schuleffektivität, insbesondere das Modell von Creemers und Kyriakides, auf die Implementation des Zentralabiturs, bestimmt die bildungspolitische Initiierung der Reform der Prüfungsorganisation und damit einhergehende Intentionen und Anreize für eine Verbesserung der Leistungen den Kontext der Schule (ggf. mit einer Erhöhung des Inputs, z. B. durch zusätzliche Ressourcen). Einzelschulische Veränderungen innerschulischer Strukturen, Abläufe, Handlungen oder Verständigungsprozesse über das Zentralabitur schlagen sich in der Gestaltung des Unterrichts nieder, beispielsweise durch eine Ausrichtung des Unterrichts an den Anforderungen und inhaltlichen Schwerpunktthemen des Zentralabiturs (teaching to the test, test preparation; Kapitel 2). Auf Seiten der Schülerinnen und Schüler könnte sich dies, auch im Sinne des Modells von Bishop (1999), auf veränderbare Variablen

wie Anstrengungen und Motivation auswirken, welche in interdependentem Zusammenhang mit Leistungen stehen (Output, Outcome).

Verbindung von Schulentwicklung und Schuleffektivität

Im Fokus von Schulentwicklung und Schuleffektivität stehen „gute" Schulen, wenn auch aus unterschiedlichen Blickwinkeln, vor verschiedenen Theoriehintergründen und mit unterschiedlichen Methoden betrachtet (für eine Übersicht der Differenzen siehe z. B. Bischof, 2017, S. 47f.; Creemers, Kyriakides, & Antoniou, 2013a, S. 114f.; Feldhoff et al., 2015, S. 65ff.; Scheerens, 1992, S. 103f.).

> There are intrinsic differences between the school effectiveness tradition, which ultimately is a programme for research with a focus on theory and explanation, and the school improvement tradition, which is a programme for innovation focusing on change and problem solving in educational practice (Creemers et al., 2013a, S. 114).

Trotz der Differenzen gibt es immer wieder Überlegungen, die beiden Theorie- und Forschungsstränge zu verbinden. Reynolds spricht von einem „integrated effectiveness/improvement paradigm" (Reynolds, 2005, S. 24), Wissinger (2007, S. 111) zufolge ist die Schuleffektivitätsforschung bereits zur Schulentwicklungsforschung geworden. Die Schuleffektivitätsforschung bietet Erkenntnisse über die Wirkung und Wirksamkeit einzelner Faktoren im Bildungsbereich und wie bessere Lernergebnisse erzielt werden können. Sie gibt eine Orientierung bzw. Richtung für Entwicklungen (innerhalb) der einzelnen Schulen. Verbunden damit ist die Erwartung, „dass eine Evidenzbasierung von Schulentwicklung zu Effektivitätszuwächsen im Bildungssystem führen kann" (Feldhoff et al., 2015, S. 82; auch Hopkins et al., 1994; Scheerens, 1992).

> Schulentwicklungsforschung braucht Schulwirksamkeitsforschung, weil Bedingungsfaktoren zu analysieren und wirksame Faktoren zu identifizieren sind, um Qualitätsverbesserungen, Prozesse, Voraussetzungen und Bedingungen untersuchen zu können (Holtappels, 2010b, S. 27).

Um die Erkenntnisse der Schuleffektivitätsforschung enger mit der Schulentwicklung der Einzelschulen verbinden zu können, entwickelten Creemers, Kyriakides und Kollegen basierend auf ihrem *dynamic model of educational effectiveness* den *dynamic approach to school improvement (DASI)*. Grundannahme

ist, dass sich jede Schule, unabhängig vom Grad ihrer Effektivität, kontinuierlich mit dem Ziel, die Leistungen der Schülerinnen und Schüler zu steigern, entwickeln sollte. Vorgesehen ist, dass die Einzelschulen sechs Stufen durchlaufen (etwa Zielklarheit und verschiedene Arten der Evaluation), dass sich jedoch die Strategien der Schulentwicklung nach den einzelschulischen Bedingungen, Zielen, Erfahrungen und Expertise richten. Der Prozess kann von verschiedenen Akteuren auf unterschiedlichen Ebenen initiiert werden. Begleitet wird die Schule durch ein Beratungs- und Forschungsteam mit Expertise in den Bereichen Schuleffektivität und Schulentwicklungspraxis. (Creemers et al., 2013a, S. 117ff.; Creemers et al., 2013b, S. 44ff.)

An dem Ansatz kritisiert Bischof (2017, S. 99f.), dass sich die Autoren darauf verlassen, dass die Schulen aufgrund eigener Einsicht den Entwicklungsprozess initiieren und umsetzen, dass sie die Empfehlungen des Beratungs- und Forschungsteams passiv annehmen statt Schulentwicklungskapazität selbstständig zu entwickeln, dass die Kommunikation zwischen Forschungsseite und Praxisseite ebenso wie der Implementationsprozess ungeklärt bleiben und dass schliesslich Hürden und Widerstände bei der Umsetzung und deren Beseitigung ausgeklammert werden.

> Creemers und Kyriakides stellen einen normativen Schulentwicklungsansatz vor, in dem die Erkenntnisse der Schulentwicklungsforschung bezüglich der Gestaltung von Veränderungsprozessen und den Hindernissen, die sich diesen entgegenstellen, keine Berücksichtigung finden (Bischof, 2017, S. 100).

In eine ähnliche Richtung weist die Kritik von Feldhoff et al. (2015, S. 81f.), dass sowohl Faktoren, welche die Schulentwicklungskapazität der Schule und die Initiierung von Veränderungen beeinflussen, als auch die Rolle schulischer Akteure und deren Handlungswissen nicht beachtet werden. Offen bleibt, welche Massnahmen in welcher Reihenfolge abzuleiten sind, um den Output der Schülerinnen und Schüler zu verbessern.

Abschliessend sei darauf verwiesen, dass nicht nur Schulentwicklung(sforschung) und Schuleffektivität(sforschung), sondern auch Educational Governance untereinander verbunden sind (Kapitel 3.3; zu Gemeinsamkeiten und Unterschieden siehe Wissinger, 2007, S. 105f.). Gemäss Brüsemeister, Altrichter und Heinrich ermöglicht Educational Governance ein „kontextualisierte[s] Verständnis von Schulentwicklung […] als einem Mehrebenenspiel" (Brüsemeister et al., 2010, S. 126).

3.6 Schlussfolgerungen aus dem theoretischen Hintergrund: Entwicklung eines Modells längerfristiger Effekte der Implementation zentraler Abiturprüfungen

Die Frage nach den Auswirkungen der Implementation zentraler Abiturprüfungen ist in den theoretischen Hintergrund der Gestaltung und Steuerung von Bildungssystemen und -prozessen unter Berücksichtigung der Mehrebenenstruktur eingebettet. Die verschiedenen theoretischen Bezüge liefern einen je eigenen Beitrag zur Analyse kurz- und längerfristiger Effekte der Einführung und Implementation zentraler Abiturprüfungen in Bremen. Die angenommenen Effekte spiegeln sich im folgenden Modell wider. Für das Zentralabitur gilt ebenso wie für high-stakes Tests: „high-stakes testing works not only as an intervention but also as an instrument to measure the outcome of the intervention" (Lee, 2008, S. 611).

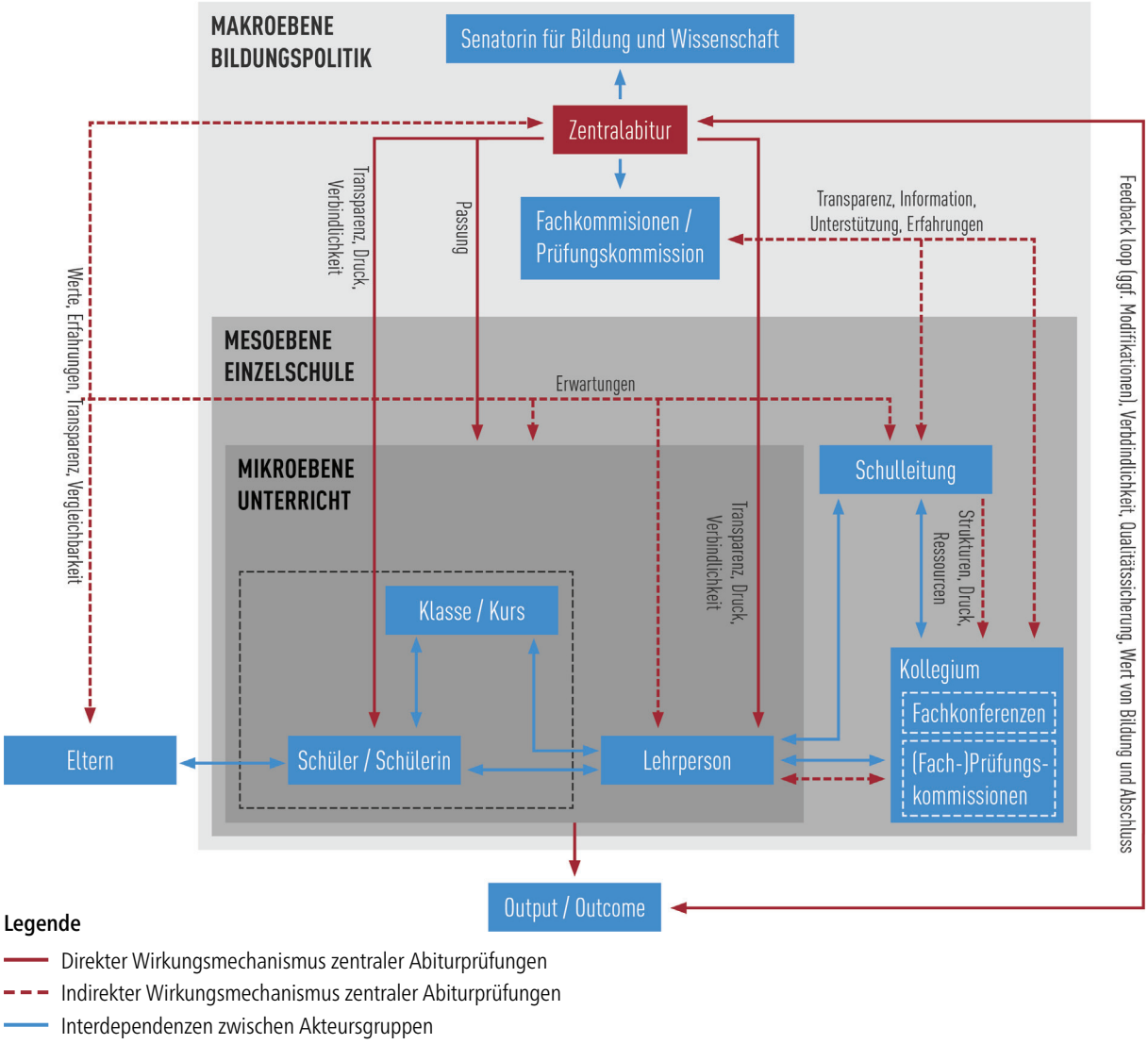


Abbildung 2: Theoretisches Rahmenmodell der kurz- und längerfristigen Effekte der Implementation zentraler Abiturprüfungen in Bremen

Um die Leistungen des Bildungssystems und die Handlungen der Akteure zu verstehen, sind die Interdependenzen und Interaktionen der verschiedenen Ebenen zu berücksichtigen (Altrichter, 2015, S. 37f.). Bezogen auf die vorliegende Arbeit bedeutet dies, dass die Initiierung der Umstellung von dezentralen zu zentralen Abiturprüfungen in Bremen durch die Bildungspolitik (Makroebene) erfolgte. Die konkrete Umsetzung und Durchführung – und letzten Endes die Implementation – liegt in der Verantwortung der einzelnen Schulen mit ihrer Eigenlogik (Mesoebene) und der der einzelnen Lehrpersonen (Mikroebene). Dafür werden die handlungsleitenden Vorgaben bezüglich des Zentralabiturs an den Kontext und die Aufgaben adaptiert bzw. rekontextualisiert (Abs et al., 2015, S. 12; Altrichter, 2015, S. 28; Fend, 2008b, S. 27). Als „Vermittler" oder „Bindeglied" zwischen der Makro- und Mesoebene fungieren die aus Lehrpersonen bestehende Prüfungskommission (plus Schulleitung), die Fachprüfungskommissionen und die Fachkonferenzen der Schulen. Auf diesen Kommunikationswegen fliessen Informationen, Wissen, Rückmeldungen, Unterstützungen, Anreize, Verbindlichkeiten, (Wert-)Vorstellungen und zwar in beide Richtungen. Der Einbezug der Lehrpersonen führt letztendlich dazu, dass eine eindeutige Trennung „zwischen Steuerungssubjekten und -objekten" (Altrichter, 2015, S. 28) nicht möglich ist.

> Aus steuerungstheoretischer Perspektive sind Lehrkräfte sowohl das Objekt als auch ein gestaltendes Subjekt von schulbezogenen Reformbemühungen, d. h. von Versuchen einer expliziten und intentionalen Gestaltung schulisch organisierter Bildungsprozesse. Sie sind zentrale Akteure, wenn es darum geht, veränderte Steuerungsimperative in eine (im Idealfall) veränderte Praxis umzusetzen und auf diese Weise mit Leben zu füllen. Insofern liegt es nahe, Lehrkräfte und ihre professionsbezogenen Einstellungen nicht nur als Effektgrößen, sondern auch als zentrale Bedingungsfaktoren von Steuerungsprozessen in und von schulischen Organisationen zu analysieren. (Koch, 2009, S. 118)

Dadurch, dass Lehrpersonen in den Prozess einbezogen werden – in Bremen etwa bei der Diskussion der Schwerpunktthemen (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4) – werden die Transparenz der Anforderungen und Erwartungen wie auch die Akzeptanz der Reform erhöht. Zugleich verringert sich das Risiko von nicht-intendiertem Verhalten, wie etwa unerwünschte Formen des teaching to the test (Kapitel 2).

> Für die Bearbeitung externer wie interner Impulse im Sinne von Entwicklungsanlässen sowie für deren wirksame Umsetzung scheint es eine Voraussetzung zu sein, dass die Lehrkräfte die Vorgaben und ihre Umsetzung als professionelles Anliegen verstehen und nicht als eine ‚von oben' oktroyierte Anforderung, mit der sie sich weder einverstanden zeigen noch identifizieren können (Baum, 2014, S. 248; Hervorhebung im Original; auch Bennewitz, 2008, S. 257).

Auf allen Ebenen und bei allen beteiligten Akteuren finden – in Abhängigkeit von je spezifischen Rahmenbedingungen, Handlungsspielräumen, Ressourcen, Überzeugungen, Interessen, Zielen und Erfahrungen, individuellen wie kollektiven Deutungen – Rekontextualisierungsprozesse statt. Diese können im Widerspruch zueinander stehen und Aushandlungen sowie Handlungskoordination erfordern (Kapitel 3.1, Kapitel 3.3). Haltungen, Meinungen und Emotionen der Akteure können die Prozesse behindern oder fördern (Bastian, 2007; Day, Elliot, & Kington, 2005; Hamilton et al., 2008, S. 34f.; Herman, 2005, S. 1f.; Kelchtermans, 2005, S. 1003; Schmidt & Datnow, 2005; van Veen et al., 2005).

Basierend auf den Ausführungen zur Implementation von Reformen und Innovationen (Kapitel 3.4) sowie zur Schulentwicklung und Schuleffektivität (Kapitel 3.5) hängen die Auswirkungen der Einführung zentraler Abiturprüfungen auf der Ebene der Einzelschule (Mesoebene) von diversen Faktoren ab. Dazu zählen Schulkultur, Schulklima, Reformbereitschaft, Expertise und Erfahrungen mit Reformen, mit Schulentwicklung und mit zentralen Abiturprüfungen, Ziele, Erwartungen, Professionalisierung, aber auch Strukturen und Aspekte der Organisation. Hinzu kommen bei der Schulleitung ebenso wie beim Kollegium, den Fachkonferenzen und den einzelnen Lehrpersonen Wissen, Ressourcen, Reformbereitschaft, Überzeugungen, Deutungsmuster, Erfahrungen und (Handlungs-)Druck, wenn auch in unterschiedlicher Konnotation. Dabei sind jeweils individuelle und kollektive Ausprägungen zu unterscheiden. Bei den einzelnen Lehrpersonen spielen zusätzlich nicht-kognitive Dispositionen, Möglichkeiten zum Austausch und Kooperationen – „als vorrangig organisatorisches Entscheidungshandeln oder […] als pädagogisch-professioneller Reflexionsraum" (Bondorf, 2013, S. 189) bzw. als Mischform beider Arten – sowie deren soziales Kapital eine Rolle. Die geringe Grösse des Bundeslandes Bremen und die im Vergleich zu anderen Bundesländern niedrige Anzahl an Schulen mit gymnasialer Oberstufe bieten für die Lehrpersonen zudem die Möglichkeit, sich als Mitglied eines „Gesamtbremer Kollegiums" zu verorten.[9] Kognitive und nicht-kognitive Dispositionen, zeitvariante und zeitinvariante Merkmale sind bei den Schülerinnen und Schülern – einzeln sowie als Klasse bzw. Kurs – von Bedeutung. Hinzukommen

[9] Dieses schulübergreifende Denken und Fühlen zeigte sich in Gesprächen mit Bremer Lehrpersonen. Inwiefern sich dies empirisch abbilden lässt, müssten weitere, ggf. qualitative, Analysen klären.

„indirekte" Erfahrungen mit dem Zentralabitur durch Kontakte zu älteren Mitschülerinnen und Mitschülern oder ehemaligen Abiturientinnen und Abiturienten.

Vor dem Hintergrund des mehrebenentheoretischen Zusammenspiels von Prozessen und Wirkungen gesellschaftlicher und schulischer (Rahmen-)Bedingungen auf unterschiedlichen Ebenen sowie von Akteuren, ihren Verantwortungen und Interaktionen lässt sich das interdependente Verhältnis von Lehrperson, Klasse, Schülerinnen und Schülern mit dem Angebot-Nutzungs-Modell von Fend (2008b, S. 21ff.) beschreiben. Kern des Modells ist die auf der Mikroebene des Unterrichts in einem ko-konstruktiven Prozess stattfindende Synchronisierung der Handlungen von Lehrpersonen (Angebotsseite) und Lernenden (Nutzungsseite), die von diversen Rahmenbedingungen des Kontextes indirekt beeinflusst sind und sich in Lernergebnissen niederschlagen. Mit der Veränderung des Kontextes durch die Umstellung von dezentralen zu zentralen Abiturprüfungen gewinnt der den Abiturprüfungen vorgelagerte Unterricht und dessen Qualität an Bedeutung, da nun eine Passung zwischen Unterricht und den allen Beteiligten unbekannten Prüfungsaufgaben hergestellt werden muss.

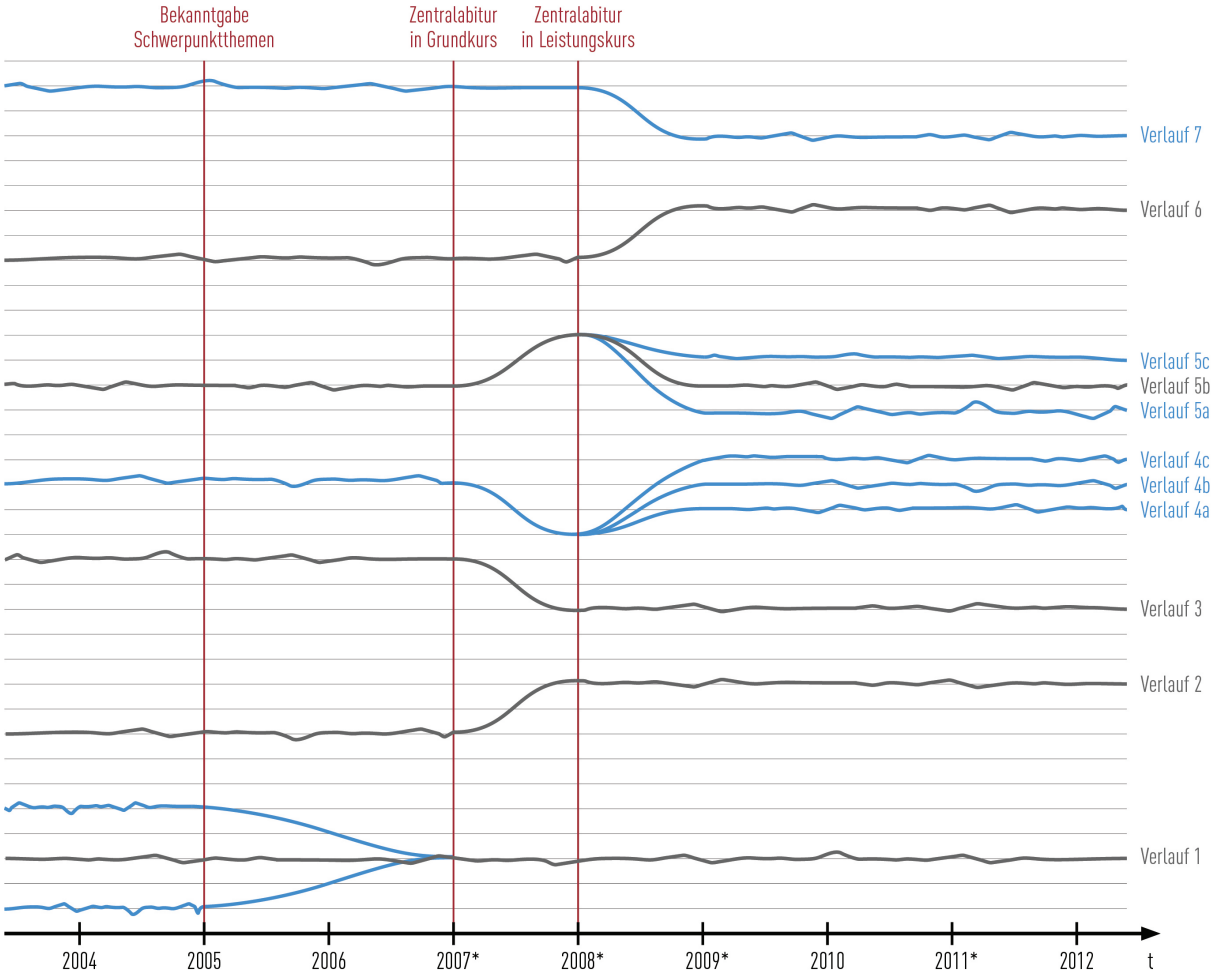
Das Angebot-Nutzungs-Modell von Fend (2008b, S. 21ff.) wurde vielfach adaptiert, modifiziert und spezifiziert, beispielsweise von Helmke (2014), Reusser und Pauli (2010), Seidel (2014) und Ditton (2000), welcher zusätzlich die Differenzierung nach intendiertem, implementiertem und erreichtem Curriculum sowie nach kurz- und längerfristigen Effekten (Output – Outcome) einführt. Den Modellen gemeinsam ist die Verflechtung und Wechselwirkung von Rahmenbedingungen des Kontextes und von Akteuren der Makro-, Meso- und Mikroebene, die sich mittelbar und unmittelbar auf den Unterricht und dessen Ergebnisse auswirken (zu Chancen von und Kritik am Angebots-Nutzungs-Modell siehe Kohler & Wacker, 2013).

Direktes Resultat des „Arbeitsbündnisses" von Lehrperson, Schülerinnen und Schülern[10] ist der Unterricht, kurzfristiges Ergebnis ist der Output und längerfristiges Ergebnis ist der Outcome, die allesamt im Zuge der Reform und der damit einhergehenden „Aufwertung" des Abiturs (Bishop & Wößmann, 2004; Piopiunik et al., 2014; Wößmann, 2003) stärker im Fokus von Schule, Lehrpersonen, Schülerinnen und Schülern, Eltern, Bildungspolitik und -verwaltung, Öffentlichkeit und Wissenschaft stehen. Sie dienen als Indikator für den Erfolg oder Misserfolg der Umstellung von dezentralen zu zentralen Abiturprüfungen

[10] Strenggenommen handelt es sich bei dieser Konstellation nicht um ein Arbeitsbündnis im Sinne von Oevermann (Oevermann, 2002, 2008). Dieses ist lediglich zwischen einer Lehrperson und einer einzelnen Schülerin bzw. einem einzelnen Schüler möglich.

und der damit verbundenen Intentionen Anforderungen zu vereinheitlichen sowie Standards, Vergleichbarkeit und Unterrichtsentwicklung zu sichern (Kapitel 2.2). Die Lern- und Prüfungsergebnisse der Schülerinnen und Schüler bilden nicht nur den Endpunkt eines Prozesses, sondern sie sind zugleich der Ausgangspunkt für Massnahmen der Qualitätssicherung. Diese Qualitätssicherung soll in Bremen unter anderem mittels des Vergleichs der Erwartungshorizonte und Korrekturhinweise zu den zentralen Prüfungen mit den korrigierten Prüfungen, der Berücksichtigung der Diskussion der schriftlichen und mündlichen Prüfungen für die Prüfungen des nächsten Jahres sowie einzelschulspezifische und landesweite Informationen zu den Ergebnissen erreicht werden (Die Senatorin für Bildung und Wissenschaft, 2013b, S. 15; Kühn, 2012, S. 37). Im Zuge des Zentralabiturs sollen die Lern- und Prüfungsergebnisse der Schülerinnen und Schüler durch Schul- und Unterrichtsentwicklung wie auch Professionalisierung der Lehrpersonen gesteigert werden (Maag Merki, 2016, S. 159f.; Maag Merki & Werner, 2013, S. 301). Ziele, Handlungen und Ergebnisse sind bei zentralen Prüfungen enger miteinander verbunden als bei dezentralen Prüfungen (Hamilton et al., 2008, S. 34f.; Herman, 2005, S. 1f.). Dies geht mit einer Veränderung der Konstellationen, Interaktionen und Interdependenzen der unterschiedlichen Akteure ebenso wie der rechtlichen, organisatorischen und kulturellen Rahmenbedingungen und Ressourcen, das heisst der Funktionslogik, einher (Benz, 2009, S. 50; Bormann & Hamborg, 2015, S. 296; Kussau & Brüsemeister, 2007, S. 28ff.). Die Umstellung von dezentraler zu zentraler Prüfungsorganisation verändert Strukturen, als deren Folge Handlungen modifiziert sowie das Verhältnis von Struktur und Handlung, von strukturierten und strukturierenden Handlungen, neu ausgehandelt werden müssen (Kapitel 3.3).

Diesbezüglich stellt sich die Frage, wieviel Zeit vergeht, bis Veränderungen in unterschiedlichen Bereichen und in den Handlungen verschiedener Akteure bzw. Akteursgruppen stattfinden, sichtbar werden und wie lange sie andauern (Gogolin et al., 2011; Hopkins et al., 1994; Thoonen et al., 2012). Um diese Frage zu beantworten, richtet die vorliegende Arbeit den Blick auf kurz- wie längerfristige Effekte der Umstellung von dezentraler zu zentraler Prüfungsorganisation und damit auf die Chronoebene. Unter Rückgriff auf das Modell von Maag Merki (2014, S. 63), erweitert durch die in Kapitel 3.2 vorgebrachte Kritik, lassen sich vier Phasen mit jeweils drei möglichen Entwicklungsrichtungen (Zunahme, Reduktion, Stagnation) unterscheiden. Der Zeitraum potentieller Effekte der Einführung und Implementation des Zentralabiturs in Bremen geht über die reine Projektlaufzeit mit dem Schwerpunkt der Datenerhebungen 2007 bis 2009 und 2011 hinaus (vierte Phase) und umfasst in stärkerem Ausmass die Zeitspanne vor der erstmaligen Durchführung zentraler Abiturprüfungen in den Grundkursen (erste Phase).



Legende

Modifikation Verlauf Maag Merki (2014)
Erweiterung Elisabeth Maué

*= Jahr der Datenerhebung

Verlauf 1: Stabilität im Zeitverlauf

Verlauf 2: Kurzfristige Zunahme (2007-2008), längerfristige Konsolidierung auf höherem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 3: Kurzfristige Reduktion (2007-2008), längerfristige Konsolidierung auf niedrigerem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 4: Kurzfristige Reduktion (2007-2008), längerfristige Konsolidierung auf
a) niedrigerem, b) höherem, c) gleichem
Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 5: Kurzfristige Zunahme (2007-2008), längerfristige Konsolidierung auf
a) niedrigerem, b) gleichem, c) höherem
Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 6: Kurzfristige Stabilität (2007-2008), verzögerte Zunahme, längerfristige Konsolidierung auf höherem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 7: Kurzfristige Stabilität (2007-2008), verzögerte Reduktion, längerfristige Konsolidierung auf niedrigerem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Zusätzlich sind bei allen Verläufen Veränderungen (Zunahme oder Reduktion) spätestens mit Bekanntgabe der Schwerpunktthemen im Jahr 2005 für die erste Durchführung zentraler Abiturprüfungen in den Grundkursen 2007 denkbar.

Abbildung 3: Weiterentwicklung des Modells der möglichen Entwicklungen/Wirkungen in Abhängigkeit der Zeit von Maag Merki (2014, S. 63)

Spätestens mit der Bekanntgabe der Schwerpunktthemen für die zentralen Abiturprüfungen (für Prüfungskommissionsmitglieder und andere noch eher) beginnt für den Grossteil der Akteure und Akteursgruppen die Auseinandersetzung mit diesen und damit die *erste Phase*. Bereits zu diesem Zeitpunkt ist mit ersten Auswirkungen zu rechnen, da, wenn auch ohne Kenntnis der Prüfungsaufgaben, der Ablauf und die Anforderungen der neuen Prüfungsform kommuniziert, verstanden sowie der Unterricht entsprechend geplant werden müssen. Hiermit dürften ein gesteigerter Zeitaufwand, Ab- und Rücksprachen, aber auch Unsicherheiten einhergehen. Darüber hinaus könnte die Reform als Eingriff „von oben" in den eigenen „Hoheitsbereich", der die eigenen Verantwortlichkeiten und Freiheiten einschränkt, empfunden werden. „Allein die Setzung von Standards und die Rückmeldung sowie ggf. Veröffentlichung von Leistungsdaten kann das Feld in rege Aktivität versetzen" (Bellmann, 2016, S. 31).

Die Umstellung kann jedoch auch ohne viel „Reibungsverluste" im Rahmen des regulären Schulalltages, sei es aufgrund entsprechender Organisation und Strukturen, sei es aufgrund von Versuchen, Bestehendes zu bewahren und möglichst wenig zu ändern, bewältigt werden (Bennewitz, 2008, S. 257; Brüsemeister, 2010, S. 14; Kussau & Brüsemeister, 2007, S. 42f.; Rogers, 2003, S. 471: „stable equilibrium").

> Allgemeiner formuliert lässt sich resümieren, dass veränderte Steuerungsverfahren dort auf Gegenliebe stoßen, wo sie am wenigsten Probleme verursachen und wo zugleich relevante Möglichkeitsstrukturen bestehen, den Anforderungen (bereits) nachzukommen (Koch, 2009, S. 134).

Die *zweite Phase* beginnt kurz vor der erstmaligen Durchführung zentraler Abiturprüfungen in den Grundkursen 2007. Lehrpersonen, Schülerinnen und Schüler haben sich zwei Jahre auf das Unbekannte vorbereitet. Kurz vor den Prüfungen dürften Druck, Stress und Unsicherheiten bezüglich der (inhaltlichen) Passung der Vorbereitung, des Ablaufs, der Prüfungsinhalte und Aufgabenformate (erneut) ansteigen.

> External exit exams are mostly aimed at providing incentives for the students, although they may also create indirect accountability pressures for teachers and schools (Woessmann et al., 2009, S. 15).

Gleichzeitig dürfte sich die intendierte Entlastung der Lehrpersonen, die nicht mehr die Prüfungsaufgaben entwickeln müssen (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4), noch nicht beim ersten Mal einstellen. Sie wird von zusätzlichem Aufwand für die Einarbeitung in die externen Korrekturkriterien sowie von Unsicherheiten und Druck überlagert. Das könnte bei Lehrpersonen das Gefühl der Kontrolle oder die Zeit, die für den Unterricht in anderen Fächern oder Altersgruppen bleibt, reduzieren und so letztlich die Zufriedenheit verringern. Denkbar ist auch, dass bei der ersten Durchführung des Zentralabiturs in den Grundkursen nicht alle Befürchtungen eintreten und sich dadurch eine eher pessimistische oder ablehnende Haltung abmildert. Selbstverständlich könnte auch das Gegenteil der Fall sein. Beeinflussen dürfte die Beurteilung des Zentralabiturs zudem die im Nachhinein sowohl top-down als auch bottom-up erfolgten Rückmeldungen *(feedback loops).*

Der letzte Punkt verweist darauf, dass mit der erstmaligen Durchführung das Thema Zentralabitur nicht abgeschlossen ist, sondern dass Rückmeldungen und Erfahrungen in Planung und Organisation des Folgejahres einfliessen *(dritte Phase)*. Zwar verfügen die Akteure nun über grösseres individuelles wie kollektives, durch Erfahrungen gewonnenes Wissen. Dessen Stabilität und Übertragbarkeit auf das folgende Jahr, auf andere mit dem Zentralabitur befasste Lehrpersonen und auf die zusätzliche Durchführung in bestimmten Leistungskursfächern bleiben jedoch fraglich. Durch die Ausweitung auf die Leistungskurse erhöht sich entsprechend der Kreis der „betroffenen" Akteure, insbesondere der Lehrpersonen, Schülerinnen und Schüler. Insofern ist eher von einer Steigerung als von einer Abnahme der Unsicherheiten auszugehen, wobei es zu Unterschieden zwischen den Grund- und Leistungskursen kommen kann. Folgt man der Argumentation, dass zentrale Prüfungen einen höheren Wert haben als dezentrale (Bishop & Wößmann, 2004; Piopiunik et al., 2014; Wößmann, 2003), müsste sich die Bedeutung dezentraler Leistungskurse und damit das Engagement der Akteure in diesen Fächern reduzieren.

Die beiden Zitate zu Beginn dieser Arbeit betonen die Bedeutung der Faktoren Zeit und Stabilität, so auch Thoonen et al. (2012, S. 448): "school-wide capacity for continuous improvement significantly develops over time". Demzufolge lässt sich erst nach einigen Jahren beurteilen *(vierte Phase)*, ob die mit der Reform verbundenen Ziele, etwa die Steigerung und Vergleichbarkeit der Leistungen von Schülerinnen und Schülern, die Entlastung der Lehrpersonen, die Reduktion des zu Anfang erhöhten Aufwandes und Stresses, erreicht wurden und werden. Fühlen sich Lehrpersonen kontrolliert und deprofessionalisiert

oder akzeptieren sie die bildungspolitische Reform? Wurden zu Beginn noch neue Handlungen später zu Routine? Ein längerer Betrachtungs- und Analysehorizont ist unabdingbar, um verzögert auftretende Auswirkungen sichtbar werden zu lassen, um einordnen zu können, ob es sich bei Erscheinungen der ersten drei Phasen um kurzfristige oder längerfristige Effekte oder Konstanten handelt und um zu verstehen, wie Handlungen verschiedener Akteure koordiniert werden. Erst dann kann beurteilt werden, ob sich das in Bremen gewählte Modell der starken Interaktionen der verschiedenen Ebenen, etwa durch den Einbezug der Lehrpersonen und durch Rückmeldungen, bewährt hat oder ob Handlungsbedarf für Modifikationen und Korrekturen besteht.

Zu bedenken ist, dass in allen Phasen nicht alle Akteure und Akteursgruppen gleichermassen betroffen bzw. involviert sind. So gibt es unter den Lehrpersonen innerhalb des Kollegiums eine grosse Varianz, abhängig davon, ob bereits ein- oder mehrfach ein Grund- oder Leistungskurs zentral geprüft wurde oder ob zentrale Abiturprüfungen aktuell oder im nächsten Schuljahr zum ersten Mal anstehen oder voraussichtlich die nächste Zeit nicht. In Anbetracht der Kritikpunkte an zentralen Abschlussprüfungen (Kapitel 2) ist jedoch davon auszugehen, dass letztlich alle Lehrpersonen mehr oder weniger, direkter oder indirekter mit dem Zentralabitur und dessen Auswirkungen in Berührung kommen.

Zusätzlich zur Zeitdimensionen (kurzfristig – langfristig) bewegen sich potenzielle Effekte in den Spannungsfeldern bzw. zwischen den Dimensionen kausal – nicht kausal, linear – nicht-linear, direkt – indirekt, generell – spezifisch, intendiert – nicht intendiert, erwünscht – unerwünscht (bzw. positiv – negativ), erwartet – unerwartet, oberflächlich – tiefgreifend, stark – schwach sowie 1. Ordnung – 2. Ordnung (Altrichter & Maag Merki, 2016a, S. 16; Baum, 2014, S. 248; Koch, 2009, S. 134f.; Reynolds, 2005, S. 13; Rogers, 2003, S. 30f.; van Ackeren, 2007, S. 191; Watanabe, 2004, S. 20ff.). Die Effekte können sich überlagern, ergänzen, in Widerspruch zueinander stehen oder sich gegenseitig nivellieren. Wie in den Ausführungen zum empirischen Hintergrund (Kapitel 4) dargelegt wird, sind vielfältige Aspekte und Effekte (der Implementation) zentraler Abschluss- und Abiturprüfungen bereits untersucht. Da dies jedoch zumeist vor einem relativ kurzen Zeithorizont und mit Querschnittsdaten geschah, richtet die vorliegende Arbeit den Blick erstens auf eine 5-Jahres-Perspektive, zweitens auf verschiedene Akteursgruppen – Lehrpersonen sowie Schülerinnen und Schüler – und drittens auf unterschiedliche Auswirkungen der Implementation des Zentralabiturs. Diese lassen sich in Anlehnung

an Fullans Unterscheidung des inneren Lernens als innerpersonelle Sinnfindung und des äusseren Lernens als Austausch und Kooperation (Fullan, 1999, S. 222) als Effekte auf die Innenperspektive, das heisst das emotionale Erleben des Zentralabiturs von Lehrpersonen sowie Schülerinnen und Schülern (Kapitel 6.3 und 6.4; Publikation 3 und Publikation 4 im Anhang), und als Effekte auf die Aussenperspektive, also die Vergleichbarkeit der Abitur- und Halbjahresnoten (Kapitel 6.1 und 6.2; Publikation 1 und Publikation 2 im Anhang), beschreiben. Zwischen diesen besteht ein zweidimensionales Wechselverhältnis: Sowohl bei den Emotionen, bei denen von einem interdependenten Verhältnis der Emotionen der Lehrpersonen mit denen der Schülerinnen und Schüler auszugehen ist (Becker, Goetz, Morger, & Ranellucci, 2014; Becker, Keller, Goetz, Frenzel, & Taxer, 2015; Collie, Shapka, & Perry, 2012; Fend, 2001; Frenzel, Goetz, Lüdtke, Pekrun, & Sutton, 2009; Hargreaves, 1998; Thiel, 2016), als auch bei der Benotung stehen die unterschiedlichen Akteursgruppen bzw. Ebenen in Zusammenhang, da Lehrpersonen die Leistungserbringung der Schülerinnen und Schüler, für die sie mit ihrem Unterricht den Grundstein gelegt haben, bewerten.

> Die Schule und die dort empirisch anzutreffende Praxis der Leistungsbeurteilung muss dabei auf der Ebene der Institution, auf der Ebene des professionellen Handelns der Lehrenden und auf der Ebene der subjektiven und individuellen Rezeption durch die Kinder und Jugendlichen selbst gesehen werden (Beutel, 2008, S. 201).

Sowohl bei Lehrpersonen als auch bei Schülerinnen und Schülern hängen Emotionen und Noten zusammen. Bei Lehrpersonen sind „emotions and values […] an important, if often overlooked and understudied, aspect of social sense-making process with respect to reform" (Spillane et al., 2002, S. 411) und damit Grundlage des professionellen Handelns, welches durch die Reform der Prüfungsorganisation modifiziert werden muss. Die Benotung gehört zum professionellen Handeln von Lehrpersonen, in das durch das Zentralabitur etwa durch die Einführung externer Korrekturkriterien eingegriffen wird.

> Das Geschäft der Leistungsbeurteilung als berufstypisches Handeln von Lehrerinnen und Lehrern ist so bedeutsam, dass eine Steigerung der Lern- und Schulqualität auch die Verbesserung des Beurteilungs- und Bewertungshandelns im Lehrerberuf einschließt, also die Professionalität des Lehrerhandelns in diesem Feld zu erhöhen ist (Beutel, 2008, S. 201).

Auf Seiten der Schülerinnen und Schüler geben Noten einerseits eine Rückmeldung über ihren Leistungsstand – im Vergleich zu ihrer Vorleistung (individuelle Bezugsnorm), zu ihren Mitschülerinnen und Mitschülern (soziale Bezugsnorm) oder zu einem vorab definierten Standard (kriteriale Bezugsnorm). Dieser Rückmeldeprozess ist mit Emotionen verbunden. Andererseits beeinflussen Emotionen die Leistung(serbringung), die anschliessend von den Lehrpersonen benotet und von den Schülerinnen und Schülern rezipiert wird.

> Beyond identity, the intensity of social interaction and judgment that students experience through the school's contexts may heighten the impact of the hours spent in school on socioemotional and academic development beyond the formal curriculum (Muller, 2015, S. 1).

4. Empirischer Hintergrund

Die empirische Basis der vorliegenden Arbeit umfasst Forschungsbefunde zu den Effekten zentraler Abschlussprüfungen, insbesondere zentraler Abiturprüfungen, und deren Implementation[11]. Im Sinne der Theorie des *washback* bzw. *backwash* geht es um den Einfluss von Prüfungen und Tests auf das Lehren und Lernen (Amengual Pizarro, 2010; Bishop, 1995; Cheng et al., 2004; Haertel, 2013; Prodromou, 1995; Scott, 2011). Da die Befundlage in Abhängigkeit vielfältiger Faktoren, etwa der in die Analysen einbezogenen Länder, Bundesländer, Jahrgangsstufen, Schulformen, Fächer, Kursniveaus oder Zeiträume, variiert, sind insbesondere Auswertungen der dieser Arbeit zugrundeliegenden Studie der Jahre 2007 bis 2009 von Bedeutung (Maag Merki, 2012c). Das Augenmerk liegt auf den Auswirkungen auf Lehrpersonen (Kapitel 4.1), Schülerinnen und Schüler (Kapitel 4.2) sowie auf schulische Prozesse und Unterricht als Prozess und Ergebnis der Interaktion beider Akteursgruppen (Kapitel 4.3).

4.1 Wirkungen zentraler Abiturprüfungen auf Lehrpersonen

Die Umstellung der Prüfungsorganisation von dezentralen zu zentralen Abiturprüfungen betrifft Lehrpersonen in besonderem Masse, da sie ihre professionellen Handlungsfelder Unterricht und Leistungsbeurteilung tangieren sowie Handlungsspielräume einschränken. Die Lehrpersonen sind gefordert, neue Strukturen aufzubauen. Dies geht in Bremen und Hessen im Zeitraum 2007 bis 2009 zunächst mit grösserem Leistungsdruck und erhöhter Unsicherheit einher, die jedoch im Laufe der Jahre abnehmen. Das Gefühl der Entlastung sowie die individuelle und kollektive Selbstwirksamkeit nehmen hingegen mit der Zeit zu (Jäger, 2012b; Oerke, 2012b; für Nordrhein-Westphalen: van Ackeren et al., 2012). Die bildungspolitisch intendierte Entlastung durch das Zentralabitur (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4) basiert auf dem Wegfall der zeitintensiven Erstellung der Abituraufgaben wie auch auf der mit der Unkenntnis der Prüfungsaufgaben einhergehenden neuen Rolle der Lehrpersonen und deren Verhältnis zu den Schülerinnen und Schülern, aber auch zu den fachlichen Inhalten (Böhm-Kasper & Weishaupt, 2002; Maag Merki, 2008; van Ackeren et al., 2012). Die Einschätzung des Schulklimas und der Arbeitsunzufriedenheit scheint von der Reform nicht direkt beeinflusst zu sein. Erfahrungen mit dem Zentralabitur verringern die Unsicherheit (zusätzliche Reduktion durch

[11] Für eine Zusammenfassung der Effekte von high-stakes Prüfungen siehe Maag Merki (2016, S. 163f.).

Kooperation unter den Lehrpersonen), die Arbeitszufriedenheit sowie die individuelle und kollektive Selbstwirksamkeit (Jäger, 2012b; Oerke, 2012b). Lehrpersonen mit geringer Berufserfahrung schätzen das Schulklima, die Arbeitszufriedenheit sowie die individuelle und kollektive Selbstwirksamkeit positiver ein als ihre Kollegen mit grösserer Berufserfahrung, die Unsicherheit gegenüber dem Zentralabitur hingegen höher. Dieses Ergebnis entspricht Befunden von Hargreaves (2005) und Kelchtermans (2005) zu Differenzen in Erfahrungen und Emotionen bezüglich Reformen je nach (Dienst-)Alter, das heisst je nach „Reformbiografie" bzw. „berufsbiografische[m; E. M.] Gedächtnis" (Bastian, 2007, S. 16f.; gegenteilige Befunde: Vähäsantanen, 2015).

Bereits im ersten Jahr 2007 haben sich die Lehrpersonen zum grossen Teil mit dem Zentralabitur auseinandergesetzt und es akzeptiert (Oerke, 2012a), womit ein wichtiger Schritt geschafft ist (Koch, 2009, S. 135). Der Fokus der Auseinandersetzung liegt dabei vorrangig auf den Leistungen und Problemen der Schülerinnen und Schüler. Die Auseinandersetzung mit dem Zentralabitur variiert mit der Zeit bezüglich Dauer und Tempo, sodass sich Subgruppen unterscheiden lassen (Oerke, 2012a). Dies schliesst an Befunde der Differenzierung verschiedener Cluster oder Typen von Lehrpersonen in Bezug auf ihre Einstellungen gegenüber der Neuen Steuerung (Bellmann, 2016; Koch, 2009), der erweiterten Schulautonomie (Zlatkin-Troitschanskaia et al., 2012), von top-down eingeführten Reformen (Bennewitz, 2008; Vähäsantanen, 2015) oder bezogen auf ihre Schulentwicklungskompetenz (Sahner, 2008; Terhart, 2010) an. Unterschiedliche Stärken und Schwächen der Cluster erfordern eine Ergänzung der verschiedenen Typen durch Kooperation für die Implementation einer Reform (Sahner, 2008, S. 270ff.). Kooperation unter Lehrpersonen ist jedoch voraussetzungsvoll, erfordert entsprechende Strukturen, Kulturen und individuelle wie kollektive Bereitschaft zur Zusammenarbeit und ist somit nicht in jedem Fall gewinnbringend (Baum, 2014; Bondorf, 2013; Fussangel, Dizinger, Böhm-Kasper, & Gräsel, 2010; Fussangel & Gräsel, 2011).

Eine Reform und damit einhergehende Änderungen von Anforderungen und Strukturen können ein Kooperationsanlass sein. Während Lehrpersonen in Nordrhein-Westfalen eine Erhöhung der Kooperationsintensität berichten (van Ackeren et al., 2012), zeigt Appius (2012) für Bremen und Hessen, dass die Häufigkeit der Kooperation im Längsschnitt von 2007 bis 2009, von einigen kurzfristigen Schwankungen abgesehen, stabil bleibt. Sie hängt vom Wunsch nach bzw. von den Einstellungen zu Kooperation

und vom Schulklima ab. Die Kooperation in Zusammenhang mit dem Abitur führt nicht zur Reduktion von Unsicherheit und Belastung. Ihre Intensität differiert in Abhängigkeit der Berufserfahrung. Auf die Bedeutung der Kooperation zwischen Lehrpersonen-Generationen und deren unterschiedliche Erfahrungen und Emotionen bezüglich Reformen verweist Hargreaves (2005).

Das theoretisch angenommene Spannungsfeld von (Er-)Neuerung und Bewahrung von Bestehendem in den Strukturen wie Handlungen (Kapitel 3.3) bestätigt sich auch empirisch, da „Deutungsmuster in der Spannung von Kontinuität und Wandel sowie Autonomie und Kontrolle die Auseinandersetzung mit Reformanforderungen strukturieren und das eigene Verhalten begründen und legitimieren" (Bennewitz, 2008, S. 254). Mit den im Zuge zentraler Prüfungen gestellten Anforderungen und ihren eigenen Überzeugungen bezüglich *best practice* befinden sich Lehrpersonen in einem weiteren Spannungsfeld (Brimijoin, 2005, S. 260; Kelchtermans, 2005; Klein, 2016). Dieser durch den Eingriff in das Handeln hervorgerufene Konflikt zwischen Innenwelt und Aussenwelt der Lehrpersonen berührt den bei der Kritik an high-stakes Prüfungen angesprochenen Punkt der Deprofessionalisierung (Kapitel 2), auf den auch Befunde zur Einführung des Zentralabiturs in Nordrhein-Westfalen verweisen (van Ackeren et al., 2012; auch Bennewitz, 2008 für die Einführung der Förderstufe in Sachsen-Anhalt; Gegenteil bei der Implementation von Bildungsstandards: Asbrand, Zeitler, & Heller, 2012).

4.2 Wirkungen zentraler Abiturprüfungen auf Schülerinnen und Schüler

Abschlussprüfungen haben für Schülerinnen und Schüler stets eine grosse Bedeutung und sind damit unabhängig von ihrer Organisation high-stakes. Dennoch empfinden Schülerinnen und Schüler in Ländern mit zentralen Abschlussprüfungen grösseren Stress und Druck (Bishop, 1999; Jürges & Schneider, 2010; Jürges et al., 2009; Pedulla et al., 2003; van Ackeren et al., 2012). Es ist sogar die Rede von einem „emotional damage" (Green et al., 2015, S. 1119). Der Vergleich von dezentraler (2007) und zentraler Prüfungsorganisation (2008, 2009) in Bremen verdeutlicht einen Anstieg der Unsicherheit gegenüber den Anforderungen im Abitur und eine Abnahme der Unsicherheit in Hessen in den Jahren nach der Einführung des zentralen Landesabiturs (2007 bis 2009). Die Unsicherheit bezüglich des Erfolgs im Abitur bleibt in Mathematik-Leistungskursen 2007 bis 2009 konstant, reduziert sich in

den Grundkursen hingegen von 2007 zu 2008 und verbleibt auf diesem Niveau (Maag Merki, 2012d). Der den Prüfungen vorgelagerte Unterricht kann zur Reduktion der Unsicherheit bezüglich der Anforderungen beitragen (Oerke, 2012b). Die Angst vor Misserfolg bleibt hingegen über die Jahre 2007 bis 2009 konstant (Maag Merki & Holmeier, 2012; ausgenommen Mathematik-Leistungskurs in Hessen: Maag Merki, 2012b). Mitunter können die durch das Zentralabitur transparenten Anforderungen bzw. Standards eine angstreduzierende Wirkung entfalten (Baumert & Watermann, 2000).

Zudem bestehen differenzielle Auswirkungen der Einführung zentraler Abiturprüfungen auf die Attributionen der Schülerinnen und Schüler in den Grundkursen in Bremen und in Hessen, nicht jedoch in den Leistungskursen. Diese Befunde variieren in Abhängigkeit des von den Schülerinnen und Schülern selbst eingeschätzten Erfolgs oder Misserfolgs im Abitur (Oerke, 2012b; Oerke, Maag Merki, Holmeier, & Jäger, 2011). Bezüglich des selbstregulierten Lernens der Abiturientinnen und Abiturienten zeigen sich ebenfalls differenzielle Effekte zwischen den Jahren 2007 bis 2009, den Bundesländern Bremen und Hessen, den Kursniveaus Grund- oder Leistungskurs und den untersuchten Fächern. Insgesamt fallen die Auswirkungen in einem normativen Sinn eher positiv aus, wenn auch in geringem Umfang (Maag Merki & Holmeier, 2012).

Eine Intention zentraler Abiturprüfungen ist die Sicherung und Steigerung von Leistungen der Schülerinnen und Schüler (Kapitel 2), wobei deren Realisation bezweifelt wird (Amrein & Berliner, 2002b, S. 58; van Ackeren & Bellenberg, 2004, S. 147). Einen Leistungsvorsprung von Schülerinnen und Schülern, die unter den Bedingungen zentraler Abschlussprüfungen lernen, weisen Bishop (1999) und Woessmann et al. (2009) sowie mit Blick auf die Sekundarstufe I in Deutschland Büchel, Jürges und Schneider (2003), Jürges et al. (2009) sowie Jürges und Schneider (2010) anhand internationaler Untersuchungen wie TIMSS und PISA nach. Dieser Befund ist jedoch nicht unumstritten. In Re-Analysen von Daten aus PISA-E-2003 finden Block, Klein, van Ackeren und Kühn (2011) zwar Unterschiede in den drei getesteten Domänen Lesekompetenz, Mathematikkompetenz und naturwissenschaftliche Kompetenz zwischen Bundesländern mit und ohne Zentralabitur, diese sind jedoch lediglich von geringer Grösse, sodass nicht von einem generellen Effekt, wohlgemerkt auf Schülerinnen und Schüler der neunten Klasse, ausgegangen werden kann. Eine Erklärung bietet möglicherweise der Befund, dass nicht die generelle Mathematik-Literacy, wie in PISA getestet, sondern spezifischeres, potentiell

prüfungsrelevantes „curricular knowledge" von zentralen Abschlussprüfungen am Ende der Sekundarstufe I positiv beeinflusst wird – und zwar bei den Schülerinnen und Schülern, denen die Prüfung zeitlich nah bevorsteht, nicht aber bei den Gymnasiastinnen und Gymnasiasten, die noch zwei oder drei Jahre Zeit bis zum Abitur haben (Jürges et al., 2009).

Ebenfalls auf Bundeslandebene analysieren Baumert und Watermann (2000) die Daten von TIMSS/III. Demnach erbringen Schülerinnen und Schüler am Ende der Sekundarstufe II zwar in Mathematik, nicht jedoch in Physik bessere Leistungen in den Leistungs- und Grundkursen in Bundesländern mit zentralen Abiturprüfungen als in Bundesländern mit dezentralen Abiturprüfungen. Diese sind vor allem auf eine Standardsicherung im unteren Leistungsbereich zurückzuführen, welche sich von 2007 bis 2009 ebenfalls in Mathematik-Grundkursen in Bremen, nicht jedoch in Hessen, findet. Im Jahr 2011 ist im Vergleich zu 2007 ebenfalls lediglich in Bremen ein Anstieg der Leistungen in den Mathematik-Grundkursen zu verzeichnen. Unter Kontrolle individueller Hintergrundmerkmale der Schülerinnen und Schüler ändern sich im Vergleich 2007 und 2009 die Leistungen in den Mathematik-Leistungskursen in Hessen kaum und in Bremen nicht. Auch im Jahr 2011 sind keine Leistungssteigerungen zu beobachten, allerdings reduziert sich in Bremen die Leistungsheterogenität. Im Fach Englisch kann lediglich der Zeitraum 2007 bis 2009 betrachtet werden, da der Englischtest in 2011 nicht eingesetzt wurde. Die Leistungen in den Englisch-Grundkursen bleiben in Bremen stabil (keine Analysen für Hessen wegen zu geringer Fallzahl). In den Englisch-Leistungskursen zeichnet sich hingegen eine leichte Steigerung der Leistungen mit Einführung zentraler Abiturprüfungen (Bremen) bzw. vom zweiten zum dritten Jahr der Durchführung (Hessen) ab. (Maag Merki, 2012a; 2016, S. 166f.)

Selbst bei ausschliesslicher Betrachtung der Sekundarstufe II variiert die Befundlage bezüglich der Steigerung der Leistungen durch zentrale Abiturprüfungen in Abhängigkeit des untersuchten Zeitraums, des Kursniveaus sowie des Faches: „there are many unknown differences in subject characteristics that might confound the estimated effects of accountability policy on student achievement" (Lee, 2008, S. 621).

4.3 Wirkungen zentraler Abiturprüfungen auf schulische Prozesse und den Unterricht

Die Auswirkungen zentraler Prüfungen auf Lehrpersonen auf der einen und Schülerinnen und Schüler auf der anderen Seite sind aufgrund der Interaktion im alltäglichen Geschäft von Schule, insbesondere im Unterricht, reziprok verbunden.

Unterrichtsgestaltung

Ziel des Unterrichts in den zwei Jahren vor den Abiturprüfungen ist eine fundierte Vorbereitung. Es besteht die Befürchtung, dass zentrale Tests mit einer Engführung und Konzentration der Unterrichtsinhalte auf prüfungsrelevante Themen und Aufgabenformate einhergehen (Kapitel 2). Dies ist ein Aspekt des *teaching to the test* innerhalb der Disziplin (Madaus et al., 2009) bzw. von *test preparation* (Koretz, 2008a) und zeigt sich in diversen Studien (z. B. Amrein & Berliner, 2002b; Barnes et al., 2000; Jürges & Schneider, 2010; Klein, 2016; Monfils et al., 2004; Roderick, Jacob, & Bryk, 2002; Gegenteil: Yeh, 2005). Obwohl das Zentralabitur im internationalen Vergleich einen geringen Standardisierungsgrad aufweist, finden sich für Bremen und Hessen 2007 bis 2009 teaching to the test-Effekte im Sinne einer Einschränkung der Themenvarianz in zentral geprüften Kursen, wobei ein Zusammenhang zu Merkmalen auf Seiten der Lehrpersonen, wie etwa deren Unsicherheit oder kollektive Selbstwirksamkeit, besteht (Jäger, 2012a; Maag Merki, 2008; Maag Merki & Holmeier, 2008; ebenso in Nordrhein-Westphalen: van Ackeren et al., 2012). Die Themenvarianz fällt auch noch im Jahr 2011 in zentral geprüften Kursen in Bremen geringer aus als in dezentral geprüften Kursen. Diese schlägt sich zwar nicht in einer höheren Abiturleistung bzw. Punktzahl in den schriftlichen Abiturprüfungen nieder, geht jedoch, ebenso wie beim dezentralen Abitur in 2007, mit einem erhöhten Fachinteresse einher (Oerke, Maag Merki, Maué, & Jäger, 2013). Doch auch beim dezentralen Abitur berichten Lehrpersonen über „Druck, dass die Themen für das Abitur gut durchgenommen werden, so dass der Handlungsspielraum für inhaltliche Vorgaben insbesondere vor dem Abitur als eingeschränkt wahrgenommen wird" (Maag Merki, 2008, S. 362). Bezüglich der Unterrichtsqualität zeigen sich in Hessen im ersten Jahr mit Zentralabitur 2007 zwei latente Klassen, die sich hinsichtlich des Grades der kognitiven Aktivierung im Unterricht in den Grundkursen unterscheiden. In beiden Klassen fallen die kognitive Aktivierung und die Unterstützung in den Leistungskursen höher aus als in den Grundkursen. In Bremen führt die schrittweise Einführung zu einer kurzzeitigen Verschiebung des Verhältnisses zwischen zentral geprüften Grundkursen und dezentral geprüften Leistungskursen, was sich in drei latenten Klassen widerspiegelt: 1. Fokus auf zentral

geprüfte Grundkurse, 2. Fokus auf dezentral geprüfte Leistungskurse, 3. Fokus auf Grund- und Leistungskurse. Die Konzentration auf die Grundkurse wird als ein teaching to the test-Effekt zwischen den Leistungsniveaus interpretiert. (Maag Merki, Klieme, & Holmeier, 2008; auch Maag Merki & Holmeier, 2008) Die Analysen liefern Hinweise, dass die Schulen, insbesondere die Lehrpersonen mit ihrer Unterrichtsgestaltung, unterschiedlich auf die Reform der Abiturprüfungen reagieren. Dies wird ebenfalls in fach- und kursniveauspezifischen Unterschieden bei der Unterstützungsqualität im Unterricht sowohl in Bremen als auch in Hessen von 2007 bis 2009 deutlich (Holmeier & Maag Merki, 2012). In beiden Ländern mediiert die (Kompetenz-)Unterstützung der Lehrpersonen im Unterricht darüber hinaus von 2007 bis 2011 den Einfluss des Zentralabiturs auf die Persistenz, schulische Selbstwirksamkeit und das Interesse der Schülerinnen und Schüler in Englisch-Leistungskursen und die schulische Selbstwirksamkeit in Mathematik-Leistungskursen (Maag Merki & Oerke, 2016). Der Blick auf Nordrhein-Westfalen offenbart, dass Lehrpersonen dort durch die Einführung des Zentralabiturs im Grossen und Ganzen wenig an ihrem Unterricht verändert haben (Kühn & Racherbäumer, 2013).

Abituraufgaben und Korrekturkriterien

Die Erstellung der Abituraufgaben und Korrekturkriterien erfolgt beim Zentralabitur durch eine externe Kommission. Lehrpersonen in Bremen und Hessen haben in den Jahren 2007 bis 2009 die Qualität der Abituraufgaben und der externen Korrekturkriterien sowie den Aufwand für die Korrektur eingeschätzt. Die Lehrpersonen halten das Anforderungsniveau sowie die inhaltliche Breite und Tiefe der Prüfungsaufgaben für angemessen. Die Qualität der Korrekturkriterien beurteilen sie eher positiv und den Aufwand für die Korrektur dezentraler wie zentraler Abiturprüfungen schätzen sie in etwa gleich ein (Appius & Holmeier, 2012; zur inhaltlichen Analyse von Abituraufgaben in Deutschland siehe Kühn, 2010 sowie für sechs europäische Länder Krüger, 2015). Eine Analyse der Aufgabenschwierigkeit des Zentralabiturs im Fach Mathematik in Nordrhein-Westfalen kommt zu dem Ergebnis, dass mit wenigen Ausnahmen keine Verzerrung und somit kein teaching to the test vorliegt (Kahnert et al., 2015). Dabei sehen die Autorinnen und der Autor in der Passung zwischen Qualifizierungsphase, Abiturvorgaben und Prüfungsinhalten eine „Verlässlichkeit der Vorgaben und der Angemessenheit der Prüfungsaufgaben aus Sicht der Schülerinnen und Schüler, aber auch aus der Perspektive der involvierten Lehrpersonen" (Kahnert et al., 2015, S. 110). Zudem messen die Mathematik-Abituraufgaben in Nordrhein-Westfalen dieselben generellen Kompetenzen wie der voruniversitäre Leistungstest aus TIMSS/III (Kahnert,

2014; TIMSS/III Leistungstest Mathematik: Klieme, 2000). Die Lehrpersonen sehen wenig Grund zur Kritik an den Abituraufgaben sowie an den Korrekturkriterien. Allerdings verdeutlichen Analysen der Abituraufgaben der Leistungskurse Mathematik und Englisch in Nordrhein-Westphalen, dass ein Teil der Aufgaben geschlechtsspezifische Schwierigkeitsgrade aufweist, das heisst, für Schüler nicht gleich leicht/schwer ist wie für Schülerinnen (Eickelmann, Kahnert, & Lorenz, 2013; Lorenz, 2016).

Den zentralen Abituraufgaben wird ein Innovationspotential, das Anregungen zur Unterrichtsentwicklung intendiert, zugeschrieben (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4). Weisen diese aus fachdidaktischer und psychometrischer Sicht eine entsprechende Qualität auf, spricht nichts dagegen, dass Lehrpersonen, ebenso wie Schülerinnen und Schüler, die Aufgaben vergangener Abiturjahrgänge zur Übung und Vorbereitung, aber auch als Vorlage und Anregung bei der Erstellung eigener Aufgaben in der Qualifizierungsphase verwenden (Barnes et al., 2000; Klein, 2016; Monfils et al., 2004).

Benotung und Vergleichbarkeit

Beim Zentralabitur steht in engem Zusammenhang mit den Abituraufgaben deren Benotung, welche die Vergleichbarkeit sichern und steigern soll (Bishop, 1995; Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4; van Ackeren et al., 2012). Die Vergleichbarkeit von Noten verweist auf die Verwendung der kriterialen Bezugsnorm bei summativer Leistungsbewertung (statt sozialer oder individueller Bezugsnorm). Eine externe Kommission erstellt Korrekturkriterien für die dezentrale Korrektur durch Lehrpersonen. Diese sehen jedoch einen gewissen Handlungsspielraum vor. In Bremen und Hessen wenden die Lehrpersonen laut eigener Aussage von 2007 bis 2009 am häufigsten die kriteriale Bezugsnorm an und steigern deren Verwendung in diesem Zeitraum (Holmeier, 2012a, 2013; keine Veränderung in Nordrhein-Westfalen: van Ackeren et al., 2012). Es zeigen sich einzelne Differenzen je nach Fachgruppe, Kooperationshäufigkeit, realisierter Themenvarianz, Berufserfahrung sowie Geschlecht der Lehrpersonen. Hiervon unterscheidet sich mit wenigen Ausnahmen die Sicht der Schülerinnen und Schüler in den Leistungs- und Grundkursen in Biologie, Deutsch, Englisch und Mathematik: Sie stellen grösstenteils keine Veränderung in der Anwendungshäufigkeit der kriterialen Bezugsnorm fest (Holmeier, 2012a, 2013). Akzeptanz und Anwendung der kriterialen Bezugsnorm sowie Transparenz der Anwendung liessen sich durch entsprechende Informationen und Rückmeldungen aller Beteiligten erhöhen (Barnes et al., 2000, S. 645).

Nationale wie internationale Befunde zum Zusammenhang zwischen Noten und Leistungstests verweisen auf Standardisierungseffekte zentraler Abschlussprüfungen (Haptonstall, 2010; Kahnert, 2014; Maag Merki & Holmeier, 2015; Maaz, Baeriswyl, & Trautwein, 2011; Neumann, Nagy, Trautwein, & Lüdtke, 2009; Neumann, Trautwein, & Nagy, 2011; Paeplow, 2008, 2011). Zum Teil zeigen sich Unterschiede zwischen verschiedenen Subgruppen (Schildkamp, Rekers-Mombarg, & Harms, 2012; Thorsen & Cliffordson, 2012). Für die Jahre 2007 bis 2009 lassen sich in Bremen und Hessen in den Leistungs- und Grundkursen in Mathematik und Englisch mit zwei Ausnahmen keine generellen Effekte des Zentralabiturs auf die Erhöhung der Vergleichbarkeit ausmachen – weder im Sinne paralleler Entwicklungen der Punktzahl in einem externen Leistungstest und der Abiturprüfungsnoten sowie eines engeren Zusammenhanges zwischen Note und Leistungstest noch durch eine Reduktion leistungsfremder Einflüsse auf diese Noten. Vielmehr variieren die Befunde fach-, kursniveau-, schul- und bundeslandspezifisch (Holmeier, 2012b; ausführlicher: Holmeier, 2013). Im Zeitraum 2007 bis 2011 zeichnen sich in Bremen in Mathematik hingegen Standardisierungseffekte auf die Halbjahresnoten in der viersemestrigen Qualifizierungsphase vor dem Zentralabitur sowie in den Grundkursen zusätzlich auf die Note in den Abiturprüfungen ab (Maag Merki & Holmeier, 2015). Insgesamt kann auf Basis der Befundlage nicht von einem generellen Effekt des Zentralabiturs auf die Standardisierung und Vergleichbarkeit der Noten ausgegangen werden (auch van Ackeren & Klemm, 2011, S. 64).

Damit schliesst sich die Frage an, wie Veränderungen durch zentrale Abiturprüfungen initiiert werden können.

Innovationspotenzial, Schul- und Unterrichtsentwicklung

Zentrale Abschluss- bzw. Abiturprüfungen beeinflussen das Handeln unterschiedlicher Akteursgruppen. Somit können sie sowohl Auslöser und Verstärker als auch Hindernis für Unterrichts- und Schulentwicklung sein. Die Erwartungen von bildungspolitischer Seite sind deutlich formuliert und zielen auf „eine didaktische und methodische Weiterentwicklung des Unterrichts" (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4) ab. Dafür werden unter anderem die Prüfungsergebnisse genauer analysiert sowie weitere Schritte zur Sicherung der Qualität von Unterricht und (zentralen) Prüfungen unternommen (Kapitel 2.2). In Übereinstimmung mit den theoretischen Annahmen verschiedener Rekontextualisierungsprozesse und Handlungen von Akteuren (Kapitel 3) zeigt der Blick in die USA,

dass nicht von einheitlichen Reaktionen der Lehrpersonen auszugehen ist. Vielmehr unterscheiden sie sich stärker innerhalb einer Schule als zwischen den Schulen und *districts*. Letztere haben kaum Einfluss auf das Handeln der Lehrpersonen im Unterricht und darauf, wie Unterricht durch Testdaten gelenkt wird, sodass Hamilton et al. (2008) das Potenzial der *standards-based accountability* (SBA) für Unterrichtsentwicklung kritisch sehen: „SBA has not overcome the organizational challenges to coordinated instructional improvement" (Hamilton et al., 2008, S. 32). Diese Sichtweise wird gestützt durch das auf vergleichenden Analysen in Finnland, Irland und den Niederlanden basierende Ergebnis, dass zentrale Prüfungen keinen generellen Effekt auf die datengestützte Schulentwicklung aufweisen (Klein, Krüger, Kühn, & van Ackeren, 2014, S. 16). Zudem stellen Schildkamp et al. (2012) fest, dass in den Niederlanden wenig Schulleitungen und Lehrpersonen die ihnen zur Verfügung gestellten Prüfungsdaten vertiefend analysieren und Massnahmen ableiten. Der Umgang mit den Daten ist voraussetzungsvoll und hängt von verschiedenen Faktoren ab, unter anderem von der Schulorganisation und der Kooperation, aber auch von der Qualität der Rückmeldungen, deren Akzeptanz durch die Akteure, allen voran den Lehrpersonen, oder Unterstützungen bei der Interpretation und Nutzung der Daten. Weiterhin ist Datenfeedback oft mit hohen bzw. überhöhten Erwartungen bezüglich seines Potentials zur Anregung und Steuerung von Schul- und Unterrichtsentwicklung verbunden (Altrichter, Moosbrugger, & Zuber, 2016a, S. 268; Specht, 2008, S. 49). „Die Nutzung von rückgemeldeten Informationen lässt sich ganz offensichtlich nicht direkt und einheitlich in den erzielten Effekten steuern" (van Ackeren, 2007, S. 197).

In Bremen und Hessen setzen sich 2007 bis 2009 die Lehrpersonen zwar mit dem Zentralabitur auseinander, dies jedoch vorrangig mit Blick auf die Leistungen der Schülerinnen und Schüler und mit der Zeit immer weniger mit Bezug zum eigenen Unterricht und dessen Verbesserung (Oerke, 2012a). Kühn zufolge ist das den zentralen Abiturprüfungen innewohnende Innovationspotential für den Transfer in die schulische Praxis zu abstrakt.

> To sum up, the assumption that statewide exit exams are utilized to implement innovations promptly and comprehensively is altogether not evident. Furthermore we can carefully assume that the innovation potential of individual teachers seems to be higher than that of a central exam commission. (Kühn, 2011, S. 194)

Dabei ist jedoch zu bedenken, dass die einzelnen Lehrpersonen Teil eines Kollegiums und einer der Einzelschule je eigenen Schulkultur mit eigenen Werten und Aushandlungsprozessen sind, welche den Rahmen für die individuellen Handlungen bilden (van Ackeren et al., 2015, S. 105). Die empirischen Befunde verweisen zudem auf die Bedeutung der Mehrebenenstruktur mit unterschiedlichen Interdependenzen, Interaktionen und Koordination der Handlungen und Handlungsbedingungen verschiedener Akteure. Auf dieses Zusammenspiel nimmt die Einführung zentraler Abiturprüfungen nicht nur Einfluss, sondern erfordert im Rahmen der Implementation Veränderungen und Aufbau von Handlungen und Strukturen (Kapitel 3), die sich in den hier vorgestellten Analysen einmal mehr und einmal weniger deutlich zeigen.

Die bisher vorliegenden empirischen Ergebnisse deuten darauf hin, dass zentrale Elemente des ‚neuen Steuerungsmodells' keineswegs jene ‚robusten' Interventionen darstellen, die relativ unabhängig von anderen Umgebungsbedingungen zu der erhofften Qualitätsentwicklung der Schulsysteme führen (Altrichter, 2015, S. 55).

4.4 Schlussfolgerungen aus dem empirischen Hintergrund: Forschungsdesiderat

Es gibt zahlreiche empirische Ergebnisse zu unterschiedlichen, mit zentralen Abschluss- bzw. Abiturprüfungen in Zusammenhang stehenden Aspekten. Da es jedoch fraglich ist, inwieweit sich die vorrangig aus dem englischsprachigen Raum und insbesondere aus den USA stammenden Befunde zu high-stakes Tests auf die spezifisch deutschen Bedingungen des Abiturs übertragen lassen, verringert sich die Vielfalt der Studien und Forschungsergebnisse, sobald sich der Blick erstens auf den deutschen Kontext, zweitens auf das Ende der Sekundarstufe II und drittens auf den Wechsel von dezentralen zu zentralen Abiturprüfungen richtet. Die entsprechenden Befunde differieren je nach Fach, Kursniveau, Schule und Bundesland, sodass nicht von einem generellen Effekt des Zentralabiturs auf das Lehren und Lernen ausgegangen werden kann. Insofern besitzt folgendes Zitat aus dem Jahr 2009 auch heute noch Gültigkeit, obwohl seitdem etliche Ergebnisse publiziert wurden.

> Die nationale und internationale Forschung ergibt derzeit keine eindeutige und abgesicherte Befundlage zu den Effekten von zentralen Abschlussprüfungsverfahren auf schulische, unterrichtliche und individuelle Arbeitsprozesse und -ergebnisse (Klein et al., 2009, S. 618).

Einige Studien stützen sich auf Querschnittsdaten, wie etwa diejenigen zu Abituraufgaben (Eickelmann et al., 2013; Kahnert et al., 2015; Lorenz, 2016; Ausnahme: Kühn, 2010). Die Studien, die sich auf Längsschnittdaten oder auf Querschnittsdaten mit mehreren Messzeitpunkten beziehen, insbesondere die Analysen der Daten des dieser Arbeit zugrundeliegenden Forschungsprojektes, beziehen zwar mehrere Messzeitpunkte ein, umfassen mit den Jahren 2007 bis 2009 schwerpunktmässig jedoch einen eher kurzen Zeitraum (Maag Merki, 2012c). Sie ermöglichen damit lediglich eingeschränkte Aussagen zur Stabilität der Befunde.

Diesem Defizit begegnet die vorliegende Arbeit mit der Erweiterung des Untersuchungszeitraumes von 2007 bis 2011 und ermöglicht die Differenzierung einer kurzfristigen und einer längerfristigen Perspektive und damit Aussagen zur zeitlichen Stabilität und Belastbarkeit von Befunden.

5. Übergeordnete Fragestellungen und Hypothesen

Die einzelnen Beiträge behandeln jeweils unter spezifischer Fragestellung Auswirkungen der Einführung zentraler Abiturprüfungen in Bremen. Sie ergänzen sich und leisten einen je eigenen Beitrag zur Klärung der Frage, ob sich längerfristige Effekte der Implementation zentraler Abiturprüfungen nachweisen lassen. Dabei richtet sich der Blick nicht nur auf verschiedene Akteure im Mehrebenensystem Schule – Lehrpersonen sowie Schülerinnen und Schüler – sondern auch auf verschiedene Bereiche: sowohl auf das „Äussere", das in der Vergleichbarkeit von Noten sichtbar wird, als auch auf das „Innere", das heisst das emotionale Erleben der Akteure. Es ergeben sich folgende Fragestellungen:

1. Führt die Implementation zentraler Abiturprüfungen aufgrund des im Vergleich zu dezentralen Abiturprüfungen erhöhten Grades an Standardisierung bei der Bewertung in längerfristiger Perspektive (2007 bis 2011) zu einer Steigerung der Vergleichbarkeit der Abiturnoten und der Halbjahresnoten im Fach Mathematik?

Die Implementation zentraler Abiturprüfungen verfolgt das Ziel, die Vergleichbarkeit der Noten zu erhöhen (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4). Die Hypothese, dass unter den Bedingungen des Zentralabiturs die *Noten der zentralen Abiturprüfungen* stärker auf den Leistungen der Schülerinnen und Schüler und weniger auf deren Hintergrundmerkmalen basieren als bei dezentralen Abiturprüfungen, fusst auf folgenden Mechanismen. Da erstens den Lehrpersonen für die Korrektur standardisierte Erwartungshorizonte und Korrekturkriterien zur Verfügung gestellt werden, ist davon auszugehen, dass sich die Notengebung bei den zentralen Abiturprüfungen an der kriterialen Bezugsnorm und damit an den Leistungen der Schülerinnen und Schüler orientiert. Zweitens werden mit der Analyse der zentralen Abiturprüfungen seitens der Bildungspolitik bzw. Fachgremien und entsprechenden Feedbackschleifen die Bedeutung der Vergleichbarkeit kommuniziert und bei den Beteiligten ein Bewusstsein dafür geschaffen (Die Senatorin für Bildung und Wissenschaft, 2013b; Hamilton et al., 2008; Herman, 2005; Maag Merki & Holmeier, 2015; Kapitel 2.2; Kapitel 3.2). Zudem entsteht durch Analysen und Rückmeldungen zumindest eine „Teil-Öffentlichkeit" (Transparenz) und Vergleichbarkeit der Notengebung, welche für Lehrpersonen mit einem Druck zu Auseinandersetzung und Diskussion einhergehen können. Die in Bremen installierten Monitoringmechanismen bieten die Möglichkeit, über Fragen und Aspekte der Benotung ins Gespräch zu kommen und bestehende Praktiken bei Bedarf zu modifizieren.

Geschieht dies und sind die kriterialen Korrekturkriterien akzeptiert, ist ein Transfer auf die von den Lehrpersonen in den vier Semestern vor den Abiturprüfungen vergebenen *Halbjahresnoten* denkbar. Veränderungen benötigen jedoch Zeit und gegebenenfalls mehrere Abitur-Durchgänge, sodass sich Effekte sowohl bei den Noten der zentralen Abiturprüfungen als auch der Halbjahresnoten vermutlich erst in längerfristiger Perspektive (2007 bis 2011) ausmachen lassen.

2. Wie erleben die Lehrpersonen die Implementation zentraler Abiturprüfungen in längerfristiger Perspektive (2007 bis 2011)?

Insbesondere aus Ländern mit zentralen Abschlussprüfungen mit high-stakes Charakter ist bekannt, dass diese bei Lehrpersonen zu höherem Druck und Stress sowie einer grösseren Unzufriedenheit mit der Arbeit führen (z. B. Amrein & Berliner, 2002a; Bishop, 1999; Pedulla et al., 2003). Zentrale Abiturprüfungen beinhalten auch in einem low-stakes System wie in Deutschland bzw. Bremen durchaus high-stakes Elemente für verschiedene Akteure. Für Lehrpersonen bedeutet dies eine grössere Transparenz ihrer Arbeit oder die nötige Passung der Vorbereitung im Unterricht auf die unbekannten zentralen Abiturprüfungen. Hinzu kommen durch die Reform der Prüfungsorganisation hervorgerufene Unsicherheiten. Die Befundlage für Bremen weist für die Jahre 2007 bis 2009 unter anderem zunächst eine Steigerung und anschliessende Abnahme von Leistungsdruck und Unsicherheit sowie eine über die Jahre zunehmende Entlastung aus (Oerke, 2012b). Da sich mit der Zeit Erfahrungen und Sicherheiten im Umgang mit zentralen Abiturprüfungen einstellen und Lehrpersonen über neue oder modifizierte Handlungsroutinen verfügen, ist in längerfristiger Perspektive (2007 bis 2011) von einer weiteren *Reduktion negativer Emotionen* und einer *fortschreitenden Entlastung der Lehrpersonen* auszugehen, die nicht mehr durch andere durch das Zentralabitur hervorgerufene zusätzliche Aufgaben und Belastungen nivelliert wird (Maag Merki, 2008).

3. Geht die Implementation zentraler Abiturprüfungen in längerfristiger Perspektive (2007 bis 2011) bei Schülerinnen und Schülern mit erhöhten negativen Emotionen einher?

Da das Ergebnis von Abschluss- bzw. Abiturprüfungen eine wegweisende Bedeutung für den weiteren Berufs- und Lebensweg von Schülerinnen und Schülern hat, weisen diese Prüfungen, unabhängig davon, ob sie dezentral oder zentral organisiert sind, stets einen high-stakes Charakter auf. Forschungsbefunde

legen jedoch nahe, dass Schülerinnen und Schüler in Ländern mit zentralen Abschlussprüfungen grösseren Stress und Druck empfinden (Bishop, 1999; Jürges & Schneider, 2010; Jürges et al., 2009; Pedulla et al., 2003; van Ackeren et al., 2012). Entsprechend steigt mit der Einführung des Zentralabiturs in Bremen die Unsicherheit gegenüber den Anforderungen im Abitur in den Jahren 2008 und 2009, nicht jedoch die Unsicherheit bezüglich des Erfolgs im Abitur in Mathematik-Leistungskursen und die Angst vor Misserfolg (Maag Merki, 2012d; Maag Merki & Holmeier, 2012; Oerke, 2012b). Transparente Anforderungen, Standards sowie der den Prüfungen vorgelagerte Unterricht können Unsicherheiten und Ängste mindern (Baumert & Watermann, 2000; Oerke, 2012b). Es wird angenommen, dass sich in längerfristiger Perspektive (2007 bis 2011) die *negativen Emotionen der Schülerinnen und Schüler reduzieren*, und zwar aus folgenden Gründen: Einerseits sollten sich die vielfältigen Erfahrungen der Lehrpersonen mit dem Zentralabitur (inhaltliche Schwerpunktthemen, Durchführung der Prüfungen) in einer fundierten und Sicherheit vermittelnden Vorbereitung der Schülerinnen und Schüler im Unterricht auf die zentralen Abiturprüfungen widerspiegeln. Andererseits werden die Erfahrungen älterer Abiturientinnen und Abiturienten an die Schülerinnen und Schüler weitergegeben, sodass das Zentralabitur mit der Zeit nicht mehr das „grosse Unbekannte" ist. Zudem ist die Vorbereitung auf das Abitur aufgrund einer stetig wachsenden Auswahl an Materialien, wie etwa zentrale Abituraufgaben vergangener Jahre, auch ausserhalb des Unterrichts möglich (van Ackeren et al., 2012).

Eine weitere Frage widmet sich dem Umstand, dass sich Veränderungen möglicherweise nicht auf alle Akteure gleichermassen auswirken (Altrichter & Maag Merki, 2016a).

4. Zeichnen sich Akteure ab, die von der Implementation zentraler Abiturprüfungen in längerfristiger Perspektive (2007 bis 2011) profitieren bzw. nicht profitieren?

Die theoretische Annahme von Rekontextualisierungsprozessen als Interpretation und Adaption von Rahmen- und Handlungsbedingungen sowohl der übergeordneten Ebene(n) als auch der eigenen Ebene (Fend, 2008a, 2008b) betont den Gestaltungsspielraum verschiedener Akteursgruppen wie einzelner Akteure im Mehrebenensystem Schule, deren Handlungen koordiniert werden müssen (z. B. Altrichter, 2015; Kapitel 3.3). Diese Annahme stützen zahlreiche Befunde, unter anderem von Seiten der Forschung zu Innovationen und Reformen, welche beispielsweise unterschiedliche Reaktionen von Schulen auf Impulse und Anforderungen von aussen oder verschiedene Einstellungen von (Gruppen

von) Lehrpersonen gegenüber einer Reform belegen (Baum, 2014; Bennewitz, 2008; Bormann, 2011; Buske, 2014; Koch, 2009; Schütze & Breidenstein, 2008). Auch bei der Auseinandersetzung mit dem Zentralabitur zeichnen sich Subgruppen von Lehrpersonen ab (Oerke, 2012a). Insofern ist davon auszugehen, dass *Lehrpersonen* je nach persönlichem und beruflichem Hintergrund unterschiedlich auf die Implementation des Zentralabiturs reagieren. Mögliche Ursachen könnten sein, dass die mit zentralen Abiturprüfungen verbundenen Ziele mit den beliefs und Werten einiger Lehrpersonen übereinstimmen, während anderen Lehrpersonen die Eingriffe in das eigene Handeln zu weit reichen und sie die Autonomie des Lehrberufs eingeschränkt sehen (Stichwort Deprofessionalisierung; Bellmann, 2016; Bennewitz, 2008; Black et al., 2011; Bormann, 2012; Good et al., 2010; van Ackeren et al., 2012). Demnach würden vorrangig die Lehrpersonen von der Implementation zentraler Abiturprüfungen profitieren, deren persönlicher und beruflicher Hintergrund eine „reibungslose" Auseinandersetzung mit und Umsetzung der Reform der Prüfungsorganisation ermöglicht und nicht im Gegensatz zu den Intentionen und Rahmenbedingungen des Zentralabiturs steht.

Auf Seiten der *Schülerinnen und Schüle*r lassen sich ebenso differenzielle Effekte der Implementation des Zentralabiturs mit Rekurs zu Rekontextualisierungsprozessen (Fend, 2008a, 2008b) annehmen. Auch die Schülerinnen und Schüler müssen die Vorgaben und Rahmenbedingungen des Zentralabiturs interpretieren, an ihren Kontext adaptieren und etwa durch eine Intensivierung der Vorbereitung ihre Handlungen entsprechend ausrichten. Ausserdem ist belegt, dass Schülerinnen und Schüler in Abhängigkeit verschiedener persönlicher Merkmale, wie etwa Leistungsstand, Attributionen von Erfolg bzw. Misserfolg oder Einschätzung des Unterrichts, Prüfungen unterschiedlich emotional erleben (z. B. Gläser-Zikuda & Fuß, 2008; Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004; Meijer, 2007; Oerke, 2012b; Oerke et al., 2011). Vom Zentralabitur nicht profitieren würden demnach Schülerinnen und Schüler, die Schwierigkeiten haben, den höheren Druck und Stress (Bishop, 1999; Jürges & Schneider, 2010; Jürges et al., 2009; Pedulla et al., 2003; van Ackeren et al., 2012) zu kompensieren und die mit der Implementation zentraler Abiturprüfungen verbundenen Intentionen zu erfüllen. Hinweise auf differenzielle Effekte liefern die vielfältigen, in Kapitel 4 dargestellten Forschungsbefunde, auf deren Basis nicht auf einen generellen Zentralabitureffekt geschlossen werden kann. Da die Ergebnisse selbst innerhalb Deutschlands nach Bundesland, Schule, Fach und Kursniveau differieren, ist anzunehmen, dass verschiedene Gruppen von Schülerinnen und Schülern in unterschiedlichen Bereichen von der Umstellung der Prüfungsorganisation profitieren bzw. nicht profitieren.

6. Zusammenfassung der vier Publikationen

Die vier Beiträge nehmen unterschiedliche Facetten der Frage längerfristiger Effekte der Implementation zentraler Abiturprüfungen in Bremen in den Blick. Der Fragestellung, ob das Zentralabitur längerfristig die Vergleichbarkeit der Noten im Fach Mathematik erhöht, widmen sich die Beiträge mit Fokus auf den Abiturnoten (Kapitel 6.1) und den Halbjahresnoten (Kapitel 6.2). Die Beurteilung des Lernfortschritts von Schülerinnen und Schülern und damit die Vergabe von Noten sind Handlungen von Lehrpersonen, die einerseits in Relation zu ihren eigenen Emotionen stehen und andererseits Emotionen auf Seiten der Schülerinnen und Schüler auslösen. Insofern geht ein Beitrag auf die längerfristige Entwicklung der Emotionen der Lehrpersonen in Zusammenhang mit dem Zentralabitur ein (Kapitel 6.3; Fragestellung 2) und ein weiterer Beitrag auf die Emotionen der Schülerinnen und Schüler (Kapitel 6.4; Fragestellung 3).

6.1 Vergleichbarkeit der Abiturnoten in Mathematik

Maué, E. (2013). Vergleichbarkeit von Abiturnoten – eine Fiktion? Längerfristige Effekte der Implementation zentraler Abiturprüfungen in Bremen. In J. Asdonk, S. U. Kuhnen, & P. Bornkessel (Hrsg.). Von der Schule zur Hochschule. Analysen, Konzeptionen und Gestaltungsperspektiven des Übergangs (S. 114-128). Münster: Waxmann.

Der Beitrag richtet seinen Fokus auf die Frage, ob die Implementation zentraler Abiturprüfungen längerfristig zur Erhöhung der Vergleichbarkeit der Note in der schriftlichen Abiturprüfung im Mathematik-Leistungskurs beiträgt.

Theoretischer und empirischer Hintergrund

Vor dem *theoretischen Hintergrund* der Bedeutung des Abiturs sowie der Abiturnoten stützt sich die theoretische Rahmung einerseits auf die Forderung nach Objektivität, Reliabilität und Validität von Noten (Ingenkamp, 1972; Lintorf, 2012) sowie mögliche Einflussfaktoren auf Noten (Ditton, 2007; Ingenkamp, 2005). Andererseits wirken sich Anstrengungen, etwa durch die Einführung von zentralen Abschlussprüfungen oder Bildungsstandards die Standardisierung und Vergleichbarkeit von Leistungen und Noten zu erhöhen, aus (z. B. Oelkers & Reusser, 2008).

Empirisch zeigt sich, dass Noten von diversen Faktoren beeinflusst sind. So differieren Noten nicht nur aufgrund unterschiedlicher Rahmenbedingungen zwischen Fächern, Klassen bzw. Kursen, Schulen, Schultypen und Bundesländern (z. B. Baumert & Watermann, 2000; Hochweber, 2010; Neumann et al., 2009), sondern auch aufgrund individueller Merkmale der Schülerinnen und Schüler (z. B. Bornkessel & Kuhnen, 2011; Büchel et al., 2003; Maaz et al., 2011). Dennoch lassen sich Standardisierungseffekte zentraler Abiturprüfungen erkennen (Neumann et al., 2009; Neumann et al., 2011). Bezüglich der Auswirkungen der Implementation zentraler Abiturprüfungen kann Holmeier (2012b) für die Jahre 2007 bis 2009 unterschiedliche Entwicklungen für Mathematik-Leistungskurse in Bremen und Hessen aufzeigen. Für Bremen ist festzuhalten, dass der Zusammenhang zwischen individueller Leistung und Note sowie der Einfluss des Geschlechts auf die Note (schlechtere Noten für Schüler bei gleicher Leistung wie Schülerinnen) unabhängig von der Prüfungsform über die Jahre konstant bleiben. Der ungünstige Einfluss eines ausländischen Geburtslandes auf die Note nimmt bis 2009 ab. Im ersten Jahr zentraler Abiturprüfungen 2008 erhalten Schülerinnen und Schüler mit mehr Büchern zuhause bei gleicher Leistung eine bessere Note. Dieser Effekt reduziert sich jedoch 2009 wieder.

Forschungsdesiderat, Fragestellungen und Hypothesen

Der vorliegende Beitrag ergänzt sowohl die auf Querschnittsdaten beruhenden bisherigen Befunde zu Einflüssen von organisatorischen und individuellen Merkmalen sowie von Standardisierungsprozessen auf Noten als auch die lediglich kurzfristigen (2007-2009) Analysen des Wechsels von dezentralen zu zentralen Abiturprüfungen von Holmeier (2012b), indem er die längerfristige Entwicklung von 2007 bis 2011 in den Blick nimmt.

Die *Fragestellungen* fokussieren erstens darauf, ob längerfristig die Entwicklungen der Note in der schriftlichen Abiturprüfung im Mathematik-Leistungskurs und der Leistung im Mathematiktest parallel verlaufen und in welchem Zusammenhang sie stehen. In Anbetracht der Tatsache, dass Veränderungen Zeit und Erfahrungen benötigen (Maag Merki, 2012b), werden längerfristige Auswirkungen der erhöhten Standardisierung von Prüfung und Beurteilung erwartet. Diese sollten sich längerfristig in einer Angleichung der beiden Entwicklungen (Hypothese 1) sowie einer Verstärkung des Zusammenhangs zwischen der Leistung und der Note (Hypothese 2) ausdrücken.

Zweitens geht der Beitrag der Frage nach, inwiefern zentrale Abiturprüfungen im Stande sind, Einflüsse leistungsfremder Faktoren, in diesem Fall des Geschlechts sowie der sozialen und kulturellen Herkunft der Schülerinnen und Schüler, auf die Note in der schriftlichen Abiturprüfung im Mathematik-Leistungskurs zu kompensieren. Die Annahme, dass Noten vorrangig individuelle Leistungen der Schülerinnen und Schüler widerspiegeln (Hypothese 3) und sich deren Effekt aufgrund von Standardisierungsprozessen steigert (Hypothese 4), stützt sich auf bisherige Befunde (Holmeier, 2012b; Maaz et al., 2011; Neumann et al., 2009; Neumann et al., 2011). Weiterhin sollte sich basierend auf den Ergebnissen von Holmeier (2012b) der Einfluss leistungsfremder Faktoren auf die Note mit der Zeit weiter reduzieren (Hypothese 5).

Methodik

Für die Analysen stehen die *Daten* von Schülerinnen und Schülern aus Bremen der Jahre 2007, 2008 und 2011 zur Verfügung. Pro Schule wird jeweils ein Mathematik-Leistungskurs (mit mindestens fünf Personen) einbezogen, sodass sich insgesamt n = 158 (2007), n = 157 (2008) und n = 196 (2011) Schülerinnen und Schüler aus elf Schulen (Kurse: n = 33) in der Stichprobe befinden.

Die *Auswertungsstrategien* umfassen neben deskriptiven Analysen Korrelationen zwischen der Note im schriftlichen Abitur im Mathematik-Leistungskurs und dem Ergebnis im Mathematiktest. Unterschiede zwischen den Jahren dieser beiden Variablen werden mit Varianzanalysen bzw. mittels Fishers Z-Transformation auf Signifikanz (Korrelationen) geprüft. Effektstärken geben das Ausmass dieser Differenzen an (Cohen, 1988). Abschliessend wird in Analogie zu Holmeier (2012b) ein Intercepts-and-Slopes-as-Outcomes-Modell mit HLM 6.06 (Raudenbush, Bryk, & Congdon, 2004) mit der Note im Mathematik-Abitur als abhängige Variable und dem Ergebnis im Mathematiktest, dem Geschlecht, dem Geburtsland und der Bücheranzahl im Elternhaus als unabhängige Variablen auf Level 1 (Individualebene) sowie der aggregierten Leistung im Mathematiktest und Dummy-Variablen für 2008 und 2011 auf Level 2 (Kursebene) berechnet. Einzig der Mathematiktest wird auf Level 1 und 2 zentriert (grand mean). Da 2007 das Referenzjahr bildet, zeigen signifikante Interaktionseffekte zwischen den Variablen auf Level 1 und den Jahres-Dummy-Variablen auf Level 2 Veränderungen über die Zeit an.

Befunde und deren Diskussion

Weder die Note im schriftlichen Mathematik-Abitur noch die Leistung im Mathematiktest ändern sich in längerfristiger Perspektive. Einzig das erste Jahr mit zentralen Abiturprüfungen 2008 nimmt eine Sonderposition ein, jedoch entgegen der Annahme einer parallelen Entwicklung, sodass Hypothese 1 insgesamt abzulehnen ist. Gleiches gilt für Hypothese 2, die von einer Verstärkung der Korrelation zwischen diesen beiden Variablen ausgegangen ist. Die Korrelationen bleiben über die Jahre ebenfalls konstant. Weiterhin bestehende Spielräume bei der Benotung könnten trotz zentraler Korrekturkriterien ausbleibende Standardisierungswirkungen des Zentralabiturs begründen. Subjektive Persönlichkeits- und Begabungstheorien der Lehrpersonen könnten ebenfalls zum Tragen kommen (Ditton, 2007). Die Mehrebenenanalysen bestätigen bisherige empirische Befunde (Holmeier, 2012b; Maaz et al., 2011) sowie Hypothese 3, dass die individuelle Leistung einen grossen Einfluss auf die Note hat. Allerdings steigert er sich entgegen der Annahme nicht längerfristig seit Einführung des Zentralabiturs (Hypothese 4). Dass die Noten weder bei dezentralen noch bei zentralen Abiturprüfungen in Abhängigkeit des Geschlechts und des familiären Bildungshintergrundes der Schülerinnen und Schüler differieren und der nachteilige Effekt eines ausländischen Geburtslandes längerfristig fortbesteht, ist konträr zur erwarteten Verringerung des Einflusses leistungsfremder Faktoren auf die Note (Hypothese 5). Einzig in kurzfristiger Perspektive (2007-2008) treten Unterschiede zwischen dezentralen und zentralen Abiturprüfungen auf. In längerfristiger Perspektive (2007-2011) zeigen sich hingegen keine Auswirkungen der erhöhten Standardisierung der Abiturprüfungen und deren Benotung auf die Vergleichbarkeit der Abiturnoten. Die Ergebnisse für Bremen reihen sich damit in die Befunde in anderen Bundesländern ein (z. B. Trautwein, Köller, Lehmann, & Lüdtke, 2007).

6.2 Vergleichbarkeit der Halbjahresnoten in Mathematik

Maué, E. (2016). Achievement—and what else? The standardisation of semester grades due to the implementation of state-wide exit examinations. Studies in Educational Evaluation, 51, 42-54.

Da sich die Abiturdurchschnittsnote sowohl aus den Noten in den Abiturprüfungen als auch den Noten der vier vorhergehenden Halbjahre zusammensetzt, nimmt dieser Beitrag letztere in den Blick.

Ausgehend von Überlegungen zu einer höheren Standardisierung der Noten der Abiturprüfungen aufgrund der Implementation des Zentralabiturs, wird mittels eines Vergleichs der vier Halbjahresnoten der Schülerinnen und Schüler in Mathematik-Leistungskursen des Jahres 2007 (dezentrale Prüfungen) mit denen des Jahres 2011 (zentrale Prüfungen) untersucht, ob sich dieses längerfristig ebenfalls auf die Halbjahresnoten auswirkt und deren Vergleichbarkeit steigert.

Theoretischer und empirischer Hintergrund

Auf Basis unterschiedlicher Theorierichtungen können (längerfristige) Effekte der Implementation des Zentralabiturs auf die Halbjahresnoten angenommen werden. Obwohl das Zentralabitur in Bremen lediglich für Schülerinnen und Schüler einen high-stakes Charakter besitzt, empfinden auch Lehrpersonen einen erhöhten Stress und Druck (Bishop, 1995; Woessmann et al., 2009). Aufgrund ihres professionellen Selbstverständnisses haben sie das Gefühl, für die Ergebnisse der Schule verantwortlich zu sein (Klinger & Rogers, 2011). Diese professionelle Verantwortung von Lehrpersonen in Kombination mit dem zentralen Abiturprüfungen zugeschriebenen Innovationspotenzial (Haertel, 2013; Kühn, 2012) könnte einen Ausgangspunkt für Diskussionen und Reflektionen über das Lehren und Lernen bilden. Da ein Ziel des Zentralabiturs die Erhöhung, Standardisierung und Vergleichbarkeit von Noten ist (Kapitel 2), müssen nicht nur die Noten in den Abiturprüfungen, sondern auch in den vier Halbjahren zuvor auf den Prüfstand.

Dass die Implementation neuer Tests mit intendierten und nicht-intendierten Auswirkungen auf Lehren und Lernen, beteiligte Akteure, Curriculum und das gesamte System einhergeht, steht im Mittelpunkt der theoretischen und empirischen Beschreibung von *washback* (Bishop, 1995; Cheng et al., 2004; Haertel, 2013; Prodromou, 1995). Das bedeutet im vorliegenden Fall, dass Lehrpersonen, Schülerinnen und Schüler innerhalb und ausserhalb des Unterrichts auf die Vorbereitung auf das Zentralabitur fokussieren. Hierfür ziehen sie beispielsweise nicht nur die Aufgaben, sondern auch die Korrekturkriterien vergangener Jahre heran und Lehrpersonen modifizieren diese für ihre eigenen Prüfungen (Amengual Pizarro, 2010; Black et al., 2011; van Ackeren et al., 2012).

Weiterhin wird auf die *standards-based accountability theory of actio*n und den *feedback loop* (Hamilton et al., 2008; Kapitel 3.2) rekurriert. Demnach hätte die Implementation des Zentralabiturs nicht nur einen

unmittelbaren Einfluss vor der Durchführung, sondern auch einen mittelbaren (Maag Merki, 2014; auch Goldberg & Rosswell, 2000). Bezüglich der Standardisierung der Abiturnoten spielt zudem das spezielle Monitoringsystem in Bremen eine Rolle (Die Senatorin für Bildung und Wissenschaft, 2013b; Kühn, 2012; Kapitel 2.2). Empirische Befunde belegen, dass eigene und kollektive Erfahrungen mit Prüfungen, Korrekturen, deren Kriterien und externem Feedback für die Reflektion und den kollegialen Austausch über Noten und Benotung, für die Standardisierung von Noten durch die Anwendung von Korrekturkriterien auch bei Halbjahresnoten sowie für Diskussionen über das Lehren und Lernen von Bedeutung sind (Black et al., 2011; Goldberg & Roswell, 2010; zum Zusammenhang von beruflicher Erfahrung und Benotungsbias siehe Hofer, 2015).

Da Lehrpersonen die Halbjahresnoten vergeben, werden zusätzlich Theorien und Forschungsbefunde zu deren *beliefs*, Stereotypen und Erwartungen einbezogen. *Beliefs* dienen dabei als Filter bei der Interpretation von Reformen sowie als Rahmen und Richtlinie für anstehende Aufgaben (Fives & Buehl, 2012). Stereotypen und Erwartungen können hingegen direktere Auswirkungen auf die Beurteilungen von Schülerinnen und Schüler haben. Deren Leistungen sind zum einen von ihrem Hintergrund (z. B. demographische Merkmale, frühere Leistungen und Noten), zum anderen von dessen Wahrnehmungen seitens der Lehrpersonen beeinflusst (selbsterfüllende Prophezeiung). Die Wirkung der Wahrnehmungen der Lehrpersonen kann ihrerseits vom Hintergrund der Schülerinnen und Schüler moderiert sein (Jussim, Eccles, & Madon, 1996). Stereotypen und Erwartungen von Lehrpersonen können ihre Wahrnehmungen bzw. die selbsterfüllenden Prophezeiungen verzerren (Jussim et al., 1996; van Ewijk, 2011), was sich in unterschiedlichem Verhalten gegenüber Schülerinnen und Schülern niederschlagen kann (Rosenthal & Jacobson, 1992; Tenenbaum & Ruck, 2007).

Dass sich die Erwartungen von Lehrpersonen auf die Noten und Testleistungen von Schülerinnen und Schülern auswirken, ist mittlerweile *empirisch* gut belegt (Friedrich, Flunger, Nagengast, Jonkmann, & Trautwein, 2015; Trouilloud, Sarrazin, Martinek, & Guillet, 2002). Ebenso, dass die Erwartungen in Abhängigkeit von Merkmalen der Schülerinnen und Schüler, meist Geschlecht, sozialer und ethnischer Hintergrund, differieren und in unterschiedlicher Behandlung seitens der Lehrpersonen resultieren (Lazarides & Watt, 2015; Rubie-Davies, Hattie, & Hamilton, 2006; Tenenbaum & Ruck, 2007; van Ewijk, 2011; zur Genauigkeit der Wahrnehmungen von Lehrpersonen: Jussim et al., 1996).

Darüber hinaus gehen neben der Leistung der Schülerinnen und Schüler nicht-kognitive Faktoren wie Motivation, Selbstkonzepte, soziale Aspekte und Verhalten in die Notengebung ein (Bowers, 2011; Hochweber, 2010; Klapp Lekholm & Cliffordson, 2009; Trouilloud et al., 2002). Forschungsbefunde belegen meist einen Vorteil von Schülerinnen gegenüber Schülern bei gleicher oder schwächerer Leistung (Cappellari, Lucifora, & Pozzoli, 2012; Klapp Lekholm & Cliffordson, 2009; Maaz et al., 2011). Bezüglich des Einflusses des Migrationshintergrunds auf die Noten reicht die Befundlage von (institutioneller) Diskriminierung bis zur Bevorzugung (Gomolla & Radtke, 2009; Klieme, 2003; Klieme et al., 2010; Maaz et al., 2011; Schräpler & Weishaupt, 2013). Allerdings ist in vielen Fällen der sozioökonomische Hintergrund der Schülerinnen und Schüler von wesentlich grösserer Bedeutung (Bornkessel & Kuhnen, 2011; Maaz et al., 2011; Thorsen, 2012).

Forschungsdesiderat, Fragestellung und Hypothesen

Da bislang Studien zu längerfristigen Auswirkungen zentraler Abiturprüfungen auf die vorgelagerten Halbjahresnoten fehlen, nimmt dieser Beitrag diese Frage in den Blick. Es ist zu untersuchen, ob sich die Halbjahresnoten je nach individuellem Hintergrund der Schülerinnen und Schüler (Geschlecht, Geburtsland, Bücheranzahl im Elternhaus) unterscheiden. Auf Basis zahlreicher Studienergebnisse können bessere Noten für Schülerinnen (z. B. Klapp Lekholm & Cliffordson, 2009; Maaz et al., 2011; Resh, 2010), für in Deutschland geborene Schülerinnen und Schüler (z. B. Maaz et al., 2011; Schräpler & Weishaupt, 2013) sowie für diejenigen mit mehr Büchern im Elternhaus (z. B. Bornkessel & Kuhnen, 2011; OECD, 2012) erwartet werden (Hypothese 1).

Ausserdem soll in Analogie zum Beitrag zu den Noten in den Abiturprüfungen in Mathematik-Leistungskursen (Kapitel 6.1; Publikation 1 im Anhang) geprüft werden, ob sich bei Kontrolle der Leistung ebenfalls Differenzen zeigen und ob die Einführung des Zentralabiturs längerfristig zu einer höheren Vergleichbarkeit der Noten beiträgt. Mit Bezug zur theoretisch angenommenen und empirisch bestätigten höheren Standardisierung von zentralen im Vergleich zu dezentralen Abschluss- bzw. Abiturprüfungen (Bishop, 1995; Haptonstall, 2010; Maag Merki, 2014; Maag Merki & Holmeier, 2015; Paeplow, 2011; Reardon, Atteberry, Arshan, & Kurlaender, 2009) wird längerfristig (2007-2011) von einem gesteigerten Einfluss der Leistung auf die Halbjahresnoten und einer Reduktion der Effekte leistungsfremder Faktoren ausgegangen (Hypothese 2).

Methodik

In die Analysen können lediglich die *Daten* von Schülerinnen und Schülern aus Bremen der Jahre 2007 und 2011 eingehen. Die Halbjahresnoten der Abiturientinnen und Abiturienten des Jahres 2007 entstanden allesamt unter den Bedingungen dezentraler Abiturprüfungen, wohingegen die Halbjahresnoten der Absolventinnen und Absolventen des Jahres 2011 vollständig unter zentralen Abiturprüfungen gebildet wurden (erster „Durchlauf" ist das Jahr 2010 gewesen, in dem jedoch keine Erhebung stattfand). Die Halbjahresnoten derjenigen der Jahre 2008 und 2009 entstanden unter „Mischformen". In der Stichprobe befinden sich n = 253 (2007) und n = 338 (2011) Schülerinnen und Schüler in Mathematik-Leistungskursen aller 19 Bremer Schulen. Für die Mehrebenenanalysen werden lediglich Mathematik-Leistungskurse mit mindestens fünf Personen einbezogen, sodass sich in dieser Stichprobe insgesamt n = 180 (2007) und n = 215 (2011) Schülerinnen und Schüler aus 12 Schulen (Kurse: n = 24) befinden. Trotz der geringen Stichprobengrösse auf Level 2 können gemäss Maas und Hox (2005) Mehrebenenanalysen durchgeführt werden.

Die *Auswertungsstrategien* umfassen deskriptive Analysen sowie t-Tests zur Überprüfung von Unterschieden auf Signifikanz, sowohl zwischen Subgruppen in Abhängigkeit des individuellen Hintergrundes der Schülerinnen und Schüler als auch in den Halbjahresnoten zwischen den Jahren. Ferner geben Effektstärken das Ausmass dieser Differenzen an (Cohen, 1988). Ob sich die Korrelationen zwischen den Halbjahresnoten und dem Ergebnis im Mathematiktest zwischen 2007 und 2011 signifikant unterscheiden, prüfen Fishers Z-Transformationen. In Analogie zur Publikation zur Note im schriftlichen Abitur im Mathematik-Leistungskurs (Kapitel 6.1; Publikation 1 im Anhang) werden vier Intercepts-and-Slopes-as-Outcomes-Modelle mit HLM 6.06 (Raudenbush et al., 2004) mit den Halbjahresnoten im Mathematik-Leistungskurs als abhängige Variablen und dem Ergebnis im Mathematiktest, dem Geschlecht, dem Geburtsland und der Bücheranzahl im Elternhaus als unabhängige Variablen auf Level 1 (Individualebene) sowie der aggregierten Leistung im Mathematiktest und der Dummy-Variablen 2011 auf Level 2 (Kursebene) berechnet. Ebenfalls wird einzig der Mathematiktest auf Level 1 und 2 zentriert (grand mean). Die Daten des Jahres 2007 fungieren als Referenzjahr, sodass signifikante Interaktionseffekte zwischen den Variablen auf Level 1 und der Dummy-Variablen 2011 auf Level 2 Veränderungen über die Zeit anzeigen.

Befunde und deren Diskussion

Die Befunde differieren in Abhängigkeit davon, ob in den Analysen die Kontrolle der individuellen Leistung erfolgt, welches Hintergrundmerkmal der Schülerinnen und Schüler im Fokus steht, zwischen den Jahren 2007 und 2011 sowie zwischen den vier Halbjahresnoten. Wird nicht für die individuelle Leistung kontrolliert, erhalten im Ausland geborene Schülerinnen und Schüler 2007 durchgängig schlechtere Noten, wohingegen dies 2011 lediglich bei einer Halbjahresnote der Fall ist. Im Gegensatz dazu weitet sich der Vorteil von Abiturientinnen und Abiturientinnen mit überdurchschnittlich viel Büchern im Elternhaus von zwei Halbjahresnoten in 2007 auf alle Halbjahresnoten in 2011 aus. Lediglich in einem Halbjahr in 2007 werden Schüler schlechter bewertet als Schülerinnen, ansonsten finden sich keine geschlechtsspezifischen Unterschiede. Die erste Hypothese zu besseren Noten von Schülerinnen, in Deutschland geborenen Abiturientinnen und Abiturienten sowie Schülerinnen und Schülern mit mehr Büchern im Elternhaus bestätigt sich teilweise, vorrangig bezüglich letzter Subgruppe.

Bei gleicher Leistung bestehen mehr Differenzen zuungunsten der Schüler, die mit einer Ausnahme konstant bleiben und kein einheitliches Bild vermitteln. Ähnliches gilt für den Einfluss des Geburtslandes: Bei gleicher Leistung erhalten im Ausland geborene Schülerinnen und Schüler in allen Halbjahren 2007 schlechtere Noten. 2011 kehren sich die Effekte in zwei Halbjahren um und egalisieren die schlechtere Benotung, die beiden anderen bleiben unverändert zum Nachteil dieser Abiturientinnen und Abiturienten. Während ein hoher Buchbestand im Elternhaus in 2007 lediglich im ersten Halbjahr tendenziell von Vorteil ist, ist dies 2011 mit einer Ausnahme in allen Halbjahren der Fall. Der Einfluss der individuellen Leistung auf die Halbjahresnoten ändert sich von 2007 zu 2011 nicht, sodass die zweite Hypothese zu Rückwirkungen der erhöhten Standardisierung zentraler Abiturprüfungen auf die Halbjahresnoten insgesamt abzulehnen ist. Jedoch kann die Reduktion des nachteiligen Effektes eines ausländischen Geburtsjahres in 2011 als Standardisierungseffekt interpretiert werden.

Die unterschiedlichen empirischen Befunde werden zum einen in Relation gesetzt zu bisherigen Studienergebnissen und potentiellen Einflüssen auf Noten, wie beispielsweise das von den Lehrpersonen eingeschätzte Sozialverhalten oder deren Stereotypen bezüglich der Anstrengungsbereitschaft (Bowers, 2011; Jussim et al., 1996; Klapp Lekholm & Cliffordson, 2009; Resh, 2010). Zum anderen wird der Bezug hergestellt zu Emotionen, die mit der Benotung einhergehen (Resh, 2009), und deren Zusammenhang

mit Lernen, Lehren, Motivation, Anstrengung, Leistung und Noten (Brookhart, 1997; Bürgermeister, 2014). Ausserdem werden verschiedene Erklärungen für die geringer als angenommenen Auswirkungen zentraler Abiturprüfungen auf die Halbjahresnoten diskutiert.

6.3 Emotionales Erleben des Zentralabiturs von Lehrpersonen

Maué, E., Maag Merki, K., & Oerke, B. (2012). Emotionales Erleben des Zentralabiturs von Lehrpersonen in Bremen. Längerfristige Effekte der Implementation zentraler Abiturprüfungen. In S. Hornberg, & M. Parreira do Amaral (Hrsg.). Deregulierung im Bildungswesen (S. 109-130). Münster et al.: Waxmann.

Emotionen spielen eine entscheidende Rolle bei der Umsetzung von Reformen (Hargreaves, 2004). Dieser Beitrag nimmt, basierend auf Theorien und Forschungsbefunden zu Belastung, Beanspruchung und Stress im Lehrberuf, das emotionale Erleben Bremer Lehrpersonen in Zusammenhang mit der Implementation des Zentralabiturs sowie dessen kurz- und längerfristige Entwicklung bis 2011 in den Blick.

Theoretischer und empirischer Hintergrund

Der Beitrag basiert *theoretisch* auf dem von Rudow (1994) entwickelten Rahmenmodell zur Analyse der Belastung und Beanspruchung im Lehrberuf und dessen Ergänzung um die Perspektive der Schülerinnen und Schüler (Böhm-Kasper, 2004). Weiterhin werden Theorien zu Stress und Unsicherheit (z. B. Buchwald, 2011; Krause, Dorsenmagen, & Alexander, 2011; Munthe, 2001; Schwarzer, 2000) sowie Methoden der Erfassung von Beanspruchung und Stress berücksichtigt (Klusmann, Kunter, & Trautwein, 2009; Schaarschmidt & Kieschke, 2007).

Empirisch zeichnen sich verschiedene Ressourcen und Risiken beim Umgang mit den beruflichen Anforderungen sowie mit Reformen im Bildungsbereich ab. Hierzu gehören individuelle und kollektive Selbstwirksamkeitserwartungen, persönliche Überzeugungen, Kooperation im Kollegium oder Schulklima (Fussangel et al., 2010; Fussangel & Gräsel, 2011; Gehrmann, 2007; Hargreaves, 2004; Klusmann et al., 2009; Oerke, 2012b; Rudow, 1994; Schaarschmidt & Kieschke, 2007). Stress und Belastung wirken sich auf die Arbeitszufriedenheit aus (Bieri, 2006; Gehrmann, 2007; Jäger, 2012b) und können zu

Unsicherheiten führen (Lüsebrink, 2002; Soltau & Mienert, 2010). Aufgrund der ihnen innewohnenden Unsicherheiten können Reformen als Stressoren wirken. In Ländern mit high-stakes testing berichten die Beteiligten eine Erhöhung von Stress, Angst und Müdigkeit (Amrein & Berliner, 2002b; Bishop, 1999; Pedulla et al., 2003; Putwain, 2008; Ryan, Ryan, Arbuthnot, & Samuels, 2007).

Forschungsdesiderat, Forschungsfragen und Hypothesen

Da sich die Erforschung des emotionalen Erlebens von Lehrpersonen bei Reformen im Bildungssystem vorrangig auf Querschnittsanalysen stützt bzw. bisher längsschnittlich lediglich der Zeitraum von drei Jahren untersucht wurde, nimmt dieser Beitrag eine 5-Jahres-Perspektive ein. Neben der Frage nach dem Empfinden der Lehrpersonen im Jahr 2011 steht die Frage nach der Entwicklung seit 2007 im Fokus. Es wird angenommen, dass sich aufgrund zunehmender individueller und kollektiver Erfahrungen die Unsicherheit gegenüber dem Zentralabitur (Hypothese 1) sowie der Leistungsdruck (Hypothese 2) längerfristig reduzieren. Im Gegenzug sollte die empfundene Entlastung zunehmen (Hypothese 3). Für die Arbeitsunzufriedenheit wird aufgrund von Forschungsbefunden zu ihrer Stabilität (Gehrmann, 2007; Jäger, 2012b) davon ausgegangen, dass deren Niveau über die Jahre konstant bleibt (Hypothese 4). Hypothese 5 zur Richtung des Zusammenhangs der einzelnen Dimensionen des emotionalen Erlebens stützt sich auf die Analysen von Appius (2012).

Zu fragen ist weiterhin, welche Prädiktoren einen Beitrag zur Erklärung des Niveaus der einzelnen Dimensionen des emotionalen Erlebens im Jahr 2011 leisten. Da sich eine Kooperation beim Abitur und kollektive Selbstwirksamkeit positiv auf die Problembewältigung in Zusammenhang mit dem Zentralabitur auswirken (Oerke, 2012a), sollten die abiturbezogene Kooperation, die kollektive Selbstwirksamkeit und ein gutes Schulklima negative Emotionen verringern und das Gefühl der Entlastung stärken (Hypothese 6).

Methodik

In die Berechnungen gehen die *Daten* von Lehrpersonen aus Bremen der Jahre 2007, 2009 und 2011 ein. Die Querschnitt-Stichprobe besteht aus n = 614 Lehrpersonen im Jahr 2007, aus n = 424 in 2009 und aus n = 427 in 2011. Zusätzlich befinden sich n = 85 Lehrpersonen in der Längsschnitt-Stichprobe, d. h. von ihnen liegen Daten von 2007, 2009 und 2011 vor.

Die *Auswertungsstrategien* umfassen deskriptive Analysen der Indikatoren Unsicherheit, Leistungsdruck, Entlastung und Arbeitsunzufriedenheit und deren Korrelationen untereinander. Das Ausmass der Differenzen zwischen den Jahren geben Effektstärken (Cohen, 1988) für die Querschnitt-Stichprobe und einfaktorielle Varianzanalysen mit Messwiederholung für die Längsschnitt-Stichprobe an. Regressionsanalysen ermitteln die Effekte von Erfahrung mit dem Zentralabitur, demographischen und schulischen Merkmalen auf jede der vier Dimensionen im Jahr 2011 (Querschnitt-Stichprobe).

Befunde und deren Diskussion

Entsprechend Hypothese 1 verringert sich die Unsicherheit der Lehrpersonen von 2007 zu 2011, wobei dies vor allem auf den Zeitraum zwischen 2007 und 2009 zurückgeht. Im Jahr 2011 tragen Erfahrungen mit dem Zentralabitur, ein positives Schulklima sowie tendenziell Lehrerfahrungen und Kooperation bezüglich des Zentralabiturs zur Reduktion der Unsicherheit bei. Dies stimmt mit anderen Forschungsbefunden zum Zusammenhang von routiniertem Verhalten und Unsicherheit (Munthe, 2001; Oerke, 2012b) und in Teilen mit Hypothese 6 überein. Der Leistungsdruck nimmt, wie mit Hypothese 2 erwartet, längerfristig ab, jedoch nur in der Querschnitt-Stichprobe. Im Gegensatz zu den Befunden von Bieri (2006) mindern einzig die Erfahrungen mit dem Zentralabitur den Leistungsdruck, sodass Hypothese 6 in diesem Fall abzulehnen ist. Gleiches gilt für die Entlastung, wobei diese mit der Zeit wie angenommen zunimmt (Hypothese 3). Die Stabilität der Arbeitsunzufriedenheit (Hypothese 4) bestätigt sich im Längsschnitt (Gehrmann, 2007; Jäger, 2012b), jedoch nicht im Querschnitt, wo sie sich unerwartet reduziert. Letztere wird im Jahr 2011 durch ein positives Schulklima und die kollektive Selbstwirksamkeit begünstigt (teilweise Bestätigung von Hypothese 6; auch Appius, 2012). Damit beeinflussen andere Faktoren als Unsicherheit, Leistungsdruck und Entlastung die Arbeitsunzufriedenheit. Die Korrelationen der vier Dimensionen entsprechen den angenommenen Richtungen und fügen sich in das bekannte Bild des Zusammenhangs von Unsicherheit, Arbeitszufriedenheit, Entlastung bzw. Ermüdung und Stress ein (Appius, 2012; Rudow, 1994). Die nach fünf Jahren insgesamt verringerte Belastung und Arbeitsunzufriedenheit und gestiegene Entlastung empfinden im Jahr 2011 noch nicht alle Lehrpersonen. Dies könnte sich jedoch mit zunehmenden Erfahrungen mit dem Zentralabitur in Kombination mit Kooperationsmöglichkeiten ändern (Fussangel & Gräsel, 2011).

6.4 Emotionales Erleben des Zentralabiturs von Schülerinnen und Schülern

Maué, E. (2017). Die Implementation zentraler Abiturprüfungen und deren potentielle Auswirkungen auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg von Schülerinnen und Schülern. Zeitschrift für Pädagogik, 63(6), 803-826.

Emotionen sind, eingebettet in komplexe, reziproke Interaktionen von einer Person mit ihrer Umwelt, für das Lernen und Lehren unerlässlich. In Ergänzung zum Beitrag zum emotionalen Erleben des Zentralabiturs von Lehrpersonen (Kapitel 6.3; Publikation 3 im Anhang) zeichnet dieser Beitrag das Erleben der Schülerinnen und Schüler, genauer deren Erfolgsunsicherheit im Abitur und Angst vor Misserfolg, über einen Zeitraum von fünf Jahren nach (Kohortenvergleich). Es wird geprüft, ob sich durch die Implementation des Zentralabiturs kurz- und/oder längerfristig der Einfluss unterrichtlicher und schulischer Faktoren auf diese Emotionen verändert und ob sich differenzielle Effekte in Abhängigkeit des Leistungsniveaus der Abiturientinnen und Abiturienten ausmachen lassen.

Theoretischer und empirischer Hintergrund

Der Beitrag fusst *theoretisch* auf Modellen, welche die Emotionen einer Person, ihrerseits mehrdimensionale Konstrukte (Pekrun, 2006), in komplexe Wechselwirkungen von persönlichen Merkmalen und Faktoren der (schulischen) Umwelt einbetten. Anknüpfungspunkte ergeben sich zur *Kontroll-Wert-Theorie* (Pekrun, 2006), zum *Adaptable Learning Model* (Boekaerts, 1992), zur *Selbstbestimmungstheorie der Motivation* (Deci & Ryan, 1993) sowie zum *transactional model of test anxiety* (Zeidner, 1998; auch Schumacher, 2016). Diese Modelle und Theorien betonen unterschiedliche Aspekte der Interaktion von Individuum und schulischer Umwelt. Sie zeigen die Relevanz der Lehrpersonen, ihres Verhaltens, ihrer Emotionen sowie ihres Unterrichts für die emotionalen Reaktionen und Entwicklungen von Schülerinnen und Schülern. Entscheidend sind dabei subjektive Wahrnehmungen und Bewertungen der schulischen Umwelt (appraisals) sowie die Relevanz bestimmter Handlungen. Da zentralen Abschlussprüfungen ein höherer Wert zugeschrieben wird als dezentralen Prüfungen (Wößmann, 2003), sollte deren erfolgreiches Bestehen eine grössere Relevanz für die Lernenden, wie auch für die Lehrenden, haben. Demzufolge erstreckt sich die schulische Umwelt nicht nur auf das unmittelbare Umfeld. Rahmenbedingungen auf der Systemebene und deren Veränderungen durch Reformen, wie die Implementation des Zentralabiturs, beeinflussen ebenfalls die Emotionen der Beteiligten.

Empirisch ist Angst vor oder in Prüfungen gut belegt. Auslösende Faktoren können das Gefühl mangelnder Kontrolle, schlechter Vorbereitung, hoher Leistungserwartungen, von Unsicherheit in Bezug auf das Ergebnis, aber auch das Leistungsniveau der Klasse sein (z. B. Frenzel, Pekrun, & Goetz, 2007; Ryan et al., 2007; Schumacher, 2016; Seipp, 1990; Zeidner & Schleyer, 1998). Im Umkehrschluss können die Unterrichtsgestaltung, wie transparente Leistungsstandards, Leistungserwartungen und Bewertungskriterien, Rückmeldungen, eine gute Vorbereitung, Motivierungsfähigkeiten sowie ein positives Klassen- und Schulklima die Angst reduzieren (Frenzel et al., 2007; Gläser-Zikuda & Fuß, 2008; Hospel & Galand, 2016; Ledergerber, 2015; Reyes, Brackett, Rivers, White, & Salovey, 2012; Rost & Schermer, 1987; Zeidner, 1998; Zeidner & Schleyer, 1998). Je nach Leistungsniveau der Lernenden (Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004; Muijs et al., 2005; Roos et al., 2015; Vanlaar et al., 2016) und nach Unterrichtsgestaltung (Hugener, 2008; Seifried, 2009) differieren die Auswirkungen dieser Faktoren. Weiterhin erleben Lernende und Lehrende in Ländern mit zentralen Abschlussprüfungen höheren Druck und Stress (Bishop, 1999; Jürges et al., 2009), was sich in kurzfristiger Perspektive (2007-2009) auch in Bremer Leistungskursen für die Erfolgsunsicherheit im Abitur (Oerke, 2012b; Ausnahme Mathematik: Maag Merki, 2012b), jedoch nicht für die Angst vor Misserfolg in Mathematik zeigt (Maag Merki, 2012d). Eine gute Vorbereitung im Unterricht trägt zur Minderung der Erfolgsunsicherheit im Abitur bei (Oerke, 2012b).

Forschungsdesiderat, Fragestellung und Hypothesen

Als Forschungsdesiderate zeichnen sich das Zusammenspiel der Wahrnehmung des Unterrichts mit den Emotionen der Schülerinnen und Schüler am Ende der Gymnasialzeit sowie mögliche Veränderungen dieser Wechselwirkung unter den Bedingungen der Implementation zentraler Abiturprüfung, insbesondere in einer längerfristigen Perspektive, ab. Die erste Fragestellung fokussiert die Einschätzung von Erfolgsunsicherheit im Abitur und Angst vor Misserfolg von Abiturientinnen und Abiturienten über einen Zeitraum von fünf Jahren (2007-2011). Da zu Beginn keine kollektiven Erfahrungen der schulischen Umwelt mit dem Zentralabitur vorliegen, sollten die beiden Emotionen 2008 negativer ausfallen als 2007 und 2011, wo bereits Informationen und Erfahrungen mehrerer Durchgänge vorhanden sind.

Weiterhin ist zu fragen, ob sich durch die Einführung zentraler Abiturprüfungen die Effekte der von den Schülerinnen und Schülern eingeschätzten Unterrichtsgestaltung (Vorbereitung auf das Abitur im

Unterricht, Autonomie- und Kompetenzunterstützung, Motivierungsfähigkeit und Leistungserwartungen der Lehrperson), des Schulklimas und der Halbjahresnote 13/1 auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg verändern. Ausgehend von der Trennung von Unterricht und zentraler Abiturprüfung gewinnt der vorgelagerte Unterricht und dessen Passung zu den Prüfungen an Bedeutung (Klein, 2016; Oerke et al., 2011), was sich kurz- und längerfristig in einer Steigerung der Effekte genannter Faktoren auf die Emotionen widerspiegeln sollte.

Abschiessend werden kurz- und längerfristige differenzielle Effekte in Abhängigkeit des Leistungsniveaus der Lernenden geprüft und angenommen, und zwar sowohl beim Ausgangsniveau und den Entwicklungen der Emotionen (Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004) als auch bei den Auswirkungen der unterrichtlichen und schulischen Faktoren auf die Emotionen (Muijs et al., 2005; Seifried, 2009; Vanlaar et al., 2016).

Methodik

In die Analysen werden die querschnittlichen *Daten* von 4451 Schülerinnen und Schülern in Leistungskursen mit Wechsel des Prüfsystems (Deutsch, Fremdsprachen, Mathematik, Naturwissenschaften) einbezogen: 2007 (dezentral): n = 1368, 2008 (zentral): n = 1497 und 2011 (zentral): n = 1586. Fehlende Werte wurden mit Ausnahme der Halbjahresnote 13/1 mittels multipler Imputation ergänzt (Maag Merki & Oerke, 2012, S. 51).

Die *Auswertungsstrategien* beinhalten deskriptive Analysen (über die zehn imputierten Datensätze gepoolte Mittelwerte, gepoolte Standardabweichungen und Standardfehler der Mittelwertschätzung; Rubin, 1987) der Emotionen Erfolgsunsicherheit im Abitur und Angst vor Misserfolg, der Note im Leistungskurs im Halbjahr 13/1 (Range: 0-15 Punkte) als Indikator für die individuelle Leistung, der Skalen Vorbereitung im Unterricht auf das Abitur, Kompetenzunterstützung, Autonomieunterstützung, Motivierungsfähigkeit und Leistungserwartungen der Lehrperson als Merkmale der unterrichtlichen Lernumwelt sowie des Schulklimas als Indikator der schulischen Umwelt. Die Regressionen der Variablen auf die Jahre sowie multiple Gruppenvergleiche mit M*plus* Version 7.3 (Muthén & Muthén, 1998-2012) dienen der Prüfung auf Signifikanz der kurzfristigen (2007-2008) und längerfristigen (2007-2011) Jahresvergleiche. Mittels Fishers Z-Transformation werden Jahresunterschiede zwischen den Korrelationen der Emotionen

auf Signifikanz geprüft. Explorative Strukturgleichungsmodelle (M*plus* Version 7.3; Muthén & Muthén, 1998-2012) mit multiplen Gruppenvergleichen (2007, 2008, 2011) zur Ermittlung der Anzahl und Struktur der Faktoren führen zum Ausschluss zweier Items der Skala Erfolgsunsicherheit im Abitur aufgrund von Nebenladungen. Die Effekte von individueller Leistung, Wahrnehmung des Unterrichts und Schulklima (jeweils manifest) auf die beiden latenten Emotionen werden mit Strukturgleichungsmodellen (M*plus* Version 7.3; Muthén & Muthén, 1998-2012) modelliert. Kurz- und längerfristige Differenzen zwischen den Jahren werden mit multiplen Gruppenvergleichen mit drei Gruppen (2007, 2008, 2011) untersucht. Für alle Analysen liegt (mindestens partielle) skalare Messinvarianz vor. Schrittweise Restriktionen erstens der Regressionskoeffizienten der latenten Variablen auf die manifesten Variablen, zweitens der Kovarianzen der latenten Variablen, drittens der Korrelationen zwischen den latenten Variablen und viertens der Korrelationen zwischen den manifesten Variablen zwischen den Gruppen, das heisst über die Jahre, ohne eine signifikante Verschlechterung des Modellfits im Vergleich zum weniger restriktiven Modell zeigen an, dass sich die Koeffizienten zwischen den Jahren nicht signifikant verändern (Christ & Schlüter, 2012). Sämtliche Analysen werden darüber hinaus separat für jedes Leistungsniveau der Schülerinnen und Schüler, operationalisiert über die Note im Halbjahr 13/1 und differenziert nach Quartilen (oberes, beide mittlere, unteres), durchgeführt. Aufgrund der geringen Intraclass Correlation (ICC) in jedem Jahr für beide Emotionen und der kleinen Anzahl der Einheiten auf Level 2 mit 17 Schulen (Scherbaum & Ferreter, 2009), wird auf eine mehrebenenanalytische Auswertung verzichtet.

Befunde und deren Diskussion

Basierend auf den Befunden, dass sich bei latenter Modellierung die Erfolgsunsicherheit im Abitur kurzfristig reduziert, längerfristig jedoch keine Unterschiede zwischen den Jahren und damit Prüfungsformen vorliegen und dass die Angst vor Misserfolg über die Jahre stabil bleibt, ist die Hypothese einer kurzfristigen Verstärkung und einer längerfristigen Reduktion für beide Emotionen abzulehnen (analog 2007-2009: Maag Merki, 2012d; entgegen Jürges et al., 2009). Möglicherweise lassen sich die geringen Veränderungen zwischen den Jahren mit einem Transfereffekt der Implementation zentraler Abiturprüfungen in den Grundkursen 2007 auf die Leistungskurse erklären, sodass dort trotz dezentraler Prüfung bereits höhere Unsicherheiten herrschten.

Dass hohe Vorbereitung im Unterricht auf die Abiturprüfungen, Kompetenzunterstützung und Halbjahresnote 13/1 die Angst vor Misserfolg und zumeist auch die Erfolgsunsicherheit im Abitur reduzieren und dass hohe Leistungserwartungen der Lehrperson, Autonomieunterstützung und Motivierungsfähigkeit die Emotionen verstärken, wird mit Bezug zu den theoretischen (Boekaerts, 1992; Deci & Ryan, 1993; Zeidner, 1998) und empirischen Bezügen (Frenzel et al., 2007; Gläser-Zikuda & Fuß, 2008; Hospel & Galand, 2016; Ledergerber, 2015; Oerke, 2012b) zur komplexen Interrelation von Person und (schulischer) Umwelt sowie zu den Auswirkungen der Unterrichtsgestaltung auf die Emotionen von Lernenden diskutiert. Eine Steigerung der Effekte unterrichtlicher und schulischer Merkmale auf die Emotionen aufgrund der grösseren Bedeutung des vorgelagerten Unterrichts (Klein, 2016; Oerke et al., 2011) findet sich entgegen der Hypothese nicht.

Durchgängig zeigen Schülerinnen und Schüler des unteren Leistungsniveaus hypothesenkonform eine ungünstigere Einschätzung der unterrichtlichen und schulischen Dimensionen bei höherer Erfolgsunsicherheit im Abitur und Angst vor Misserfolg als diejenigen der beiden anderen Leistungsgruppen (Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004). Bezüglich der kurz- und längerfristigen Entwicklung der Emotionen finden sich zwar Unterschiede zwischen den Leistungsgruppen, jedoch bestätigt sich die Annahme ungünstigerer Entwicklungen der Emotionen bei leistungsschwächeren Lernenden lediglich für die Erfolgsunsicherheit im Abitur. Differenzielle Auswirkungen der unterrichtlichen und schulischen Umwelt je nach Leistungsniveau (Muijs et al., 2005; Vanlaar et al., 2016) lassen sich nicht einheitlich ausmachen. Zudem gibt es wenig Variation zwischen den Jahren und damit zwischen den Bedingungen dezentraler oder zentraler Abiturprüfungen.

Insgesamt kann auch für Schülerinnen und Schüler am Ende der Sekundarstufe II gezeigt werden, dass (zumindest ein Teil ihrer) Emotionen in das komplexe Geflecht aus individuellen Merkmalen der Person sowie unterrichtlichen und schulischen Faktoren eingebettet sind, die ihrerseits wiederum unter anderem durch die Emotionen der Lehrpersonen beeinflusst werden (Jennings & Greenberg, 2009). Dieses Zusammenspiel bleibt über unterschiedliche Kohorten und Rahmenbedingungen der Prüfungsorganisation hinweg relativ stabil.

7. Diskussion

Ziel der vorliegenden Arbeit ist die Analyse längerfristiger Auswirkungen der Umstellung von einer dezentralen Prüfungsorganisation hin zu zentralen Abiturprüfungen über einen Zeitraum von fünf Jahren. Zur Einordnung der Befunde bedarf es jedoch nicht nur der längerfristigen Perspektive (2007 bis 2011), sondern auch der kurzfristigen Perspektive (2007 bis 2008 bzw. 2009). Erst die Kombination beider Perspektiven ermöglicht die Abschätzung des Verlaufs der Effekte (Kapitel 3.6).

Eine ausführliche Diskussion der einzelnen empirischen Ergebnisse findet sich in den jeweiligen Beiträgen. Eine Zusammenfassung ausgewählter empirischer Befunde mit Bezug zu den übergeordneten Fragestellungen (Kapitel 7.1) ist deren theoretischer Einordnung (Kapitel 7.2) vorangestellt. Es folgen Limitationen (Kapitel 7.3) und Folgerungen für künftige Forschung zu Reformen und die Praxis (Kapitel 7.4).

7.1 Beantwortung der Forschungsfragen

Die erste übergeordnete Fragestellung bezieht sich auf die mit der Einführung zentraler Abiturprüfungen verbundene Intention einer Steigerung der Vergleichbarkeit der Noten (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4).

1. Führt die Implementation zentraler Abiturprüfungen aufgrund des im Vergleich zu dezentralen Abiturprüfungen erhöhten Grades an Standardisierung bei der Bewertung in längerfristiger Perspektive (2007 bis 2011) zu einer Steigerung der Vergleichbarkeit der Abiturnoten und der Halbjahresnoten im Fach Mathematik?

Ein Aspekt der Vergleichbarkeit von Noten betrifft den *Zusammenhang von Leistung und Note*. Ein Vergleich des Zusammenhangs zwischen der Note in der schriftlichen Abiturprüfung in Mathematik und der Leistung in einem standardisierten Mathematik-Leistungstest (Korrelation und Mehrebenenmodell) zwischen den Jahren offenbart, dass der Zusammenhang in den Jahren 2008 und 2011 in derselben

Grösse fortbesteht wie 2007. Er ist demnach seit Einführung des Zentralabiturs nicht enger geworden, was im Gegensatz zu bisherigen Befunden zu einem Standardisierungseffekt zentraler Abschlussprüfungen steht (Haptonstall, 2010; Kahnert, 2014; Maag Merki & Holmeier, 2015; Maaz et al., 2011; Neumann et al., 2009; Neumann et al., 2011; Paeplow, 2008, 2011). Bezüglich des Zusammenhangs zwischen den vier Halbjahresnoten im Mathematik-Leistungskurs und dem standardisierten Leistungstest differieren die Befunde: Während sich die Korrelationen von 2007 zu 2011 verstärken, variiert der Effekt des Leistungstests auf die Halbjahresnoten im Mehrebenenmodell zwischen 2007 und 2011 nicht.

Ein weiterer Punkt der Vergleichbarkeit bezieht sich auf den *Einfluss leistungsfremder Merkmale* der Schülerinnen und Schüler auf die Note in der schriftlichen Abiturprüfung in Mathematik bei Kontrolle der individuellen Leistung. In kurzfristiger Perspektive ergeben sich im ersten Jahr zentraler Abiturprüfungen 2008 Änderungen des Effekts des Geburtslandes und des familiären Bildungshintergrundes auf die Note, nicht jedoch des Geschlechts. Die Effekte nivellieren sich allerdings längerfristig, sodass sich kein Unterschied zwischen dezentralen Abiturprüfungen 2007 und zentralen Abiturprüfungen 2011 ausmachen lässt und der nachteilige Effekt eines ausländischen Geburtslandes fortbesteht. Die Befunde zu den Halbjahresnoten 2007 und 2011 differieren zwischen den einzelnen Halbjahren, insbesondere beim Einfluss des Geschlechts ohne konsistentes Bild (vom Fortbestand des Vorteils für Schülerinnen über den Fortbestand keiner Differenzen zwischen den Geschlechtern bis zum Ausgleich der Differenzen). Zudem nivelliert sich einerseits der 2007 in allen vier Halbjahren bestehende nachteilige Effekt eines ausländischen Geburtslandes 2011 in zwei Halbjahren. Andererseits wird 2011 der Einfluss des familiären Bildungshintergrundes in zwei Halbjahren, in denen er 2007 nicht von Bedeutung war, signifikant. Die Halbjahresnoten sind demnach stärker von leistungsfremden Merkmalen beeinflusst als die Note in der schriftlichen Abiturprüfung, was bereits andere Untersuchungen zeigten (Maaz et al., 2011; Neumann et al., 2009; Neumann et al., 2011).

Weder hinsichtlich des Zusammenhangs von Leistung und Mathematiknote in der schriftlichen Abiturprüfung sowie in den vorhergehenden vier Halbjahren noch bezüglich der Effekte leistungsfremder Merkmale auf die Noten lassen sich eindeutige Auswirkungen einer höheren Standardisierung durch zentrale Abiturprüfungen mit einheitlichen Korrekturkriterien ausmachen. Insgesamt ist nicht von einer generellen Erhöhung der Vergleichbarkeit der Noten durch die Einführung des Zentralabiturs mit entsprechenden Monitoring- und Feedbackschleifen auszugehen und diese Hypothese abzulehnen (ähnlich

für 2007-2009: Holmeier, 2012b; Holmeier, 2013). Die Befunde zu den Effekten von Geburtsland und familiärem Bildungshintergrund auf die Note im schriftlichen Mathematik-Abitur zwischen 2007 und 2008 verdeutlichen die Notwendigkeit der längerfristigen Perspektive für die Einordnung kurzfristiger Veränderungen. Dass diese Variationen keine Stabilität aufweisen und auftretende Effekte sich wieder reduzieren, wird erst mit einem längeren zeitlichen Analysefenster erkennbar.

Worauf lassen sich die vorliegenden Befunde zurückführen? Die Verwendung der zentralen Erwartungshorizonte und Korrekturkriterien sowie ein Bewusstsein für deren Bedeutung, auch aufgrund des speziellen Monitoringsystems in Bremen, wird als Wirkmechanismus angenommen (Die Senatorin für Bildung und Wissenschaft, 2013b; Hamilton et al., 2008; Herman, 2005; Maag Merki & Holmeier, 2015; Kapitel 2.2; Kapitel 3.2; Kapitel 5).

> Zudem hängt die Vergleichbarkeit von Anforderungen nicht nur von den Aufgabenstellungen, Erwartungshorizonten und Bewertungshinweisen ab, sondern auch von der *Anwendung* der Erwartungshorizonte und der Bewertungshinweise bei der Beurteilung der Leistungen von Abiturientinnen und Abiturienten (Stanat et al., 2016, S. 53; Hervorhebung im Original).

Die Lehrpersonen legen von 2007 bis 2009 verstärkt die kriteriale Bezugsnorm der Benotung zugrunde (Holmeier, 2012a, 2013) und beurteilen die Qualität der Korrekturkriterien eher positiv (Appius & Holmeier, 2012). Insofern ist zu vermuten, dass sie sich im Jahr 2011 ebenso verhalten. Nichtsdestotrotz bieten die Korrekturkriterien weiterhin Spielraum für Entscheidungen der Lehrpersonen, da sie die Schülerinnen und Schüler (mindestens) vier Halbjahre vor Durchführung der Abiturprüfungen unterrichten. In dieser Zeit entstandene, und je nach Hintergrund der Schülerinnen und Schüler variierende Eindrücke und Erwartungen können bewusst oder unbewusst in die Noten einfliessen (Cheng & Sun, 2015; Friedrich et al., 2015; Jussim et al., 1996; Lazarides & Watt, 2015; Rakoczy, Klieme, Bürgermeister, & Harks, 2008; Rubie-Davies et al., 2006; Stevens & Görgöz, 2010; Tenenbaum & Ruck, 2007; Tierney, Simon, & Charland, 2011; Trouilloud et al., 2002; van Ewijk, 2011; Weinstein, 2002). Dies könnte ebenfalls die stärkeren Einflüsse leistungsfremder Merkmale auf die Halbjahresnoten erklären, für die im Gegensatz zu zentralen Abiturprüfungen keine standardisierten Korrekturkriterien vorliegen. Eine weitere Erklärung bieten Befunde von Terhart (2008, S. 154ff.) zur schulinternen Kultur der Notengebung und Selektion,

die mittels impliziter Sozialisation von Lehrpersonen übernommen wird. Danach verheimlichen Lehrpersonen eher ihre Notengebung als dass sie über individuelle und kollektive Praktiken sprechen und diese offen aushandeln (allgemeiner: Bondorf, 2013).

> And as the practice of grading is linked to the professional and personal identity of teachers, it is also a personal practice. Teachers seem to strive for intimacy or some kind of privacy in their practice of giving marks. They try to hide the process from their colleagues, and one of the central means of hiding the process is to come up with normal, unspectacular and well-adapted results or products: Of course the marks themselves become public, but the process leading to them is kept hidden. (Terhart, 2008, S. 160)

Die Einführung zentraler Abiturprüfungen mit einem entsprechenden Monitoringsystem mit Analysen, Rückmeldungen und Diskussionen der Ergebnisse hat die Transparenz und Öffentlichkeit der Notengebung erhöht (Kapitel 2.2). Dennoch bedeutet das nicht zwangsläufig, dass Lehrpersonen über Prozesse der Benotung sprechen oder noch einen Schritt weitergehen und diese verändern. In diese Richtung weisen ebenfalls Befunde, dass Lehrpersonen zwar Aufgabeninhalte und -formate sowie Bewertungskriterien zentraler Lernstandserhebungen und Vergleichsarbeiten direkt ohne vorherige Reflexion für den eigenen Unterricht verwenden (Diemer & Kuper, 2011; Hahn, 2014). Insbesondere für den Transfer auf die von den Lehrpersonen in den vier Semestern vor den Abiturprüfungen vergebenen Halbjahresnoten bedeutet dies hohe Hürden.

Da Reformen den emotionalen Umgang von Lehrpersonen mit den Strukturen, Praktiken, Traditionen und Routinen des Lehrberufs beeinflussen (Hargreaves, 1998, S. 844), richtet die zweite Fragestellung den Fokus auf die Emotionen der Lehrpersonen.

2. Wie erleben die Lehrpersonen die Implementation zentraler Abiturprüfungen in längerfristiger Perspektive (2007 bis 2011)?

Nach fünf Jahren Zentralabitur ist insgesamt zu konstatieren, dass das emotionale Erleben zentraler Abiturprüfungen sich bei den Lehrpersonen positiv entwickelt hat. Unsicherheit, Leistungsdruck

und Arbeitsunzufriedenheit haben sich reduziert und die Entlastung ist entsprechend der Intention gewachsen (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4). Aufschlussreich ist der Zeitpunkt, zu dem sich Veränderungen beobachten lassen. Während in kurzfristiger Perspektive (2007-2009) die Unsicherheit ab- und die Entlastung zunimmt, wird die Reduktion von Leistungsdruck und Arbeitsunzufriedenheit erst in längerfristiger Perspektive sichtbar (2007-2011). Diese Befunde unterstreichen erneut die Bedeutung des längeren Untersuchungszeitraumes und schliessen an Forschungsergebnisse zu unterschiedlichen Zeitpunkten von Veränderungen in Schulentwicklungsprozessen an (Hopkins et al., 1994; Thoonen et al., 2012). Die Befunde bestätigen die Hypothese einer weiteren Reduktion negativer Emotionen und einer fortschreitenden Entlastung der Lehrpersonen. Allerdings ist bei der Unsicherheit und der Entlastung zu bedenken, dass es sich eher um eine längerfristige Stabilisierung kurzfristiger Effekte handelt. Die Ergebnisse erweitern in Übereinstimmung zu denen von Oerke (2012b; auch Jäger, 2012b; van Ackeren et al., 2012) bisherige Forschungsbefunde dahingehend, dass in Ländern mit zentralen Abschlussprüfungen Lehrpersonen nicht per se mehr Druck, Stress und Unzufriedenheit mit der Arbeit empfinden (z. B. Amrein & Berliner, 2002a; Bishop, 1999; Pedulla et al., 2003). Vielmehr müssen Kontextbedingungen wie der high- oder low-stakes Charakter der zentralen Prüfungen für die jeweilige Akteursgruppe – auch eine auf Systemebene low-stakes Prüfung kann für einzelne Akteure oder Akteursgruppen high-stakes Elemente beinhalten – und die seit der Implementation vergangene Zeit in Rechnung gestellt werden.

Insbesondere die Unsicherheit lässt sich durch Erfahrungen mit dem Zentralabitur und persönliche Lehrerfahrung, aber auch durch Schulmerkmale wie ein positives Schulklima und Kooperation in Zusammenhang mit dem Zentralabitur mindern. Ihr ist anders zu begegnen als etwa Leistungsdruck und Entlastung, die lediglich mit der Erfahrung mit dem Zentralabitur in Zusammenhang stehen. Die Arbeitsunzufriedenheit wird hingegen von einem positiven Schulklima und der kollektiven Selbstwirksamkeit abgeschwächt. Die Relation von Unsicherheit und Arbeitsunzufriedenheit zu Schulkultur und Kollegium verweist auf kollegiale Netzwerke und Interaktionen zwischen Lehrpersonen und damit auf deren soziales Kapital – eine Voraussetzung für die Implementation von Reformen und die Modifikationen von Handlungen (Penuel et al., 2012; ähnlich Bonsen, 2005; Bormann, 2011; van Ackeren et al., 2011).

Die Interpretationen von Emotionen und Erfahrungen mit einer Reform beeinflussen die professionelle Identität von Lehrpersonen, deren individuelles und kollektives *sense-making* und damit deren Einstellung

und Verhalten der Reform gegenüber sowie ihrer Nutzung (Bastian, 2007; Day et al., 2005; Kelchtermans, 2005; Maag Merki, 2012b, 2016; Schmidt & Datnow, 2005; Spillane et al., 2002; Vähäsantanen, 2015; van Veen et al., 2005; Ziegelbauer, 2015). Dieses dürfte sich entsprechend der (Educational) Governance zum einen auf die horizontalen, aber auch vertikalen Handlungskoordinationen verschiedener Akteure und Akteursgruppen mit je eigenen Handlungslogiken auswirken (Altrichter, 2015; Benz, 2009; Brüsemeister, 2010; van Ackeren et al., 2016). Zum anderen sollte sich dies im Handeln der Lehrpersonen, insbesondere in ihrer Unterrichtsgestaltung, widerspiegeln und auf diesem Weg auch die Emotionen der Schülerinnen und Schüler beeinflussen.

3. Geht die Implementation zentraler Abiturprüfungen in längerfristiger Perspektive (2007 bis 2011) bei Schülerinnen und Schülern mit erhöhten negativen Emotionen einher?

Die Hypothese der kurzfristigen Steigerung und der längerfristigen Abnahme von Erfolgsunsicherheit im Abitur sowie Angst vor Misserfolg der Schülerinnen und Schüler ist auf Grundlage der zeitlichen Stabilität der Ausprägung der Emotionen zu verwerfen (analog 2007-2009: Maag Merki, 2012d; Maag Merki & Holmeier, 2012). Einzig die Erfolgsunsicherheit im Abitur reduziert sich entgegen der Hypothese kurzfristig (2007-2008), was erneut auf die Notwendigkeit einer längerfristigen Perspektive (2007-2011) zur Einordnung der Befunde verweist. Diese stehen bisherigen Forschungsergebnissen zu erhöhtem Stress und Druck von Schülerinnen und Schülern in Ländern mit zentralen Abschlussprüfungen entgegen (Bishop, 1999; Green et al., 2015; Jürges & Schneider, 2010; Jürges et al., 2009; Pedulla et al., 2003; van Ackeren et al., 2012). Möglicherweise mindern mehrere Faktoren Unsicherheiten und Ängste: der vergleichsweise geringe Anteil der zentralen Prüfungen an der Gesamtnote (Kapitel 2.2), die neue Rolle der Lehrpersonen und damit das neue Verhältnis zwischen den Akteursgruppen (Maag Merki, 2008; van Ackeren et al., 2012; vbw – Vereinigung der Bayerischen Wirtschaft e. V., 2011), die aufgrund der bekannten Schwerpunktthemen grössere Transparenz der Prüfungsinhalte und eine darauf abgestimmte Vorbereitung als Synchronisation von Angebot- und Nutzungsseite im Sinne eines „positiven" *teaching to the test* bzw. „positiver" *test preparation* (Fend, 2008b, S. 21ff.; Firestone & Schorr, 2004, S. 2f.; Holmeier & Maag Merki, 2012; Maag Merki & Holmeier, 2008; Maag Merki et al., 2008; Natriello, 2009, S. 1108; van Ackeren & Bellenberg, 2004, S. 135f.; van Ackeren et al., 2015, S. 177). Entsprechend theoretischer Annahmen (Boekaerts, 1992; Deci & Ryan, 1993; Zeidner, 1998) und empirischer Befunde zur

Interrelation von Person und (schulischer) Umwelt sowie zu den Auswirkungen der Unterrichtsgestaltung auf die Emotionen von Lernenden (Baumert & Watermann, 2000; Frenzel et al., 2007; Gläser-Zikuda & Fuß, 2008; Hospel & Galand, 2016; Ledergerber, 2015; Maag Merki & Oerke, 2016; Meijer, 2007; Oerke, 2012b) reduzieren das Gefühl einer guten Vorbereitung im Unterricht auf die Abiturprüfungen, von Kompetenzunterstützung, eine gute Halbjahresnote 13/1 sowie ein positives Schulklima die Angst vor Misserfolg und die Erfolgsunsicherheit im Abitur (Ausnahme: Schulklima 2011). Hohe Leistungserwartungen der Lehrperson steigern hingegen die negativen Emotionen. Entgegen der Annahme wirken sich Autonomieunterstützung und Motivierungsfähigkeit auf die Erfolgsunsicherheit im Abitur nicht und auf die Angst vor Misserfolg verstärkend aus. Demnach haben zwar die Lehrpersonen und ihre Unterrichtsgestaltung einen Einfluss auf die Emotionen der Schülerinnen und Schüler, dieser nimmt jedoch nicht über die Jahre zu, trotz gesteigerter Bedeutung des den Prüfungen vorgelagerten Unterrichts durch die Implementation des Zentralabiturs (Klein, 2016; Oerke et al., 2011). Ein Grund könnte darin liegen, dass die von Jahr zu Jahr wachsende Auswahl an Materialien, beispielsweise zentrale Abituraufgaben vergangener Jahre, eine Vorbereitung auf das Abitur auch ausserhalb des Unterrichts und unabhängig von der Lehrperson ermöglicht (van Ackeren et al., 2012).

Die Befunde bieten Anhaltspunkte für differenzielle Auswirkungen auf die Akteure.

4. Zeichnen sich Akteure ab, die von der Implementation zentraler Abiturprüfungen in längerfristiger Perspektive (2007 bis 2011) profitieren bzw. nicht profitieren?

Wird zur Beantwortung dieser Frage das emotionale Erleben der *Lehrpersonen* im Jahr 2007 mit zentral geprüften Grundkursen und dezentral geprüften Leistungskursen und im Jahr 2011 mit zentral geprüften Grund- und Leistungskursen in Deutsch, Fremdsprachen, Mathematik und Naturwissenschaften herangezogen, fallen 2007 Unsicherheit, Leistungsdruck und Arbeitsunzufriedenheit höher und die Entlastung geringer aus als 2011. Insofern scheinen auf den ersten Blick alle Lehrpersonen mit der Zeit von zentralen Abiturprüfungen zu profitieren. Der günstige Einfluss individueller Erfahrung mit dem Zentralabitur auf die Emotionen verstärkt diesen Befund (Ausnahme Arbeitsunzufriedenheit). Analysen der Daten der Lehrpersonen, die in den Jahren 2007, 2009 und 2011 an der Befragung teilnahmen (Längsschnittstichprobe), offenbaren allerdings, dass sich bei dieser Personengruppe in der

längerfristigen Perspektive zwar die Unsicherheit reduziert und die Entlastung zunimmt, im Unterschied zur Querschnittstichprobe sich jedoch weder der Leistungsdruck noch die Arbeitsunzufriedenheit verringern. Die Stichprobengrösse fällt mit n = 60-85 gering aus, dennoch müssten weitere Analysen dieses emotionale Empfinden klären. Die mit einer grossen Zentralabitur-Erfahrung einhergehende Attraktivität als Ansprech- und Kooperationspartnerin bzw. -partner könnte mit zusätzlichem Aufwand und Druck verbunden sein, die Erwartungen zu erfüllen. Ähnliche Belastungen könnten auch durch eine besondere Funktionsstelle hervorgerufen werden. Zeichnen sich verschiedene Subgruppen von Lehrpersonen ab, wie es Oerke (2012a) bezüglich der Auseinandersetzung mit dem Zentralabitur zeigen konnte, ist gemäss Sahner (2008) eine Kooperation für die Implementation einer Reform gewinnbringend. Für die Frage, inwiefern das auch auf die einzelne Lehrperson bzw. unterschiedliche Subgruppen von Lehrpersonen zutrifft, wären weitere Informationen über deren längerfristige Auseinandersetzung mit dem Zentralabitur aufschlussreich, etwa (veränderte) Einstellungen, beliefs und Werte in Übereinstimmung mit den Zielen und Umsetzungen des Zentralabiturs (Baum, 2014; Bennewitz, 2008; Day et al., 2005; Hochberg & Desimone, 2010; Koch, 2009; Meredith et al., 2017; Schmidt & Datnow, 2005; Spillane et al., 2002; van Veen et al., 2005). Hierfür bieten sich als Ergänzung qualitative Interviews an. Diese können zudem tiefergehende Hinweise auf Prozesse der Rekontextualisierung (Fend, 2008a, 2008b) oder der Handlungskoordination (Altrichter, 2015; Benz, 2009) liefern.

Bei genauerer Betrachtung des Jahres 2011 fällt zudem auf, dass sich die Unsicherheit und Arbeitsunzufriedenheit durch schulische und kollegiale Aspekte wie Schulklima, Kooperation zu Fragen des Zentralabiturs (Unsicherheit) oder kollektive Selbstwirksamkeit (Arbeitsunzufriedenheit) reduzieren. Massnahmen der Teamentwicklung als Teil von Organisationsentwicklung und damit von Schulentwicklung (Rolff, 2010) bieten Anknüpfungspunkte zum Aufbau und zur Verstärkung des sozialen Kapitals von Lehrpersonen (Penuel et al., 2012; auch Jo, 2014): „professionelles Lehrerhandeln vollzieht sich stets innerhalb der Organisation Schule, wodurch sich die Möglichkeit eröffnet, Probleme nicht ausschließlich individuell zu reflektieren oder zu lösen“ (Bondorf, 2013, S. 198). Im Gegensatz dazu beeinflussen im Jahr 2011 weder schulische und kollegiale noch persönliche Merkmale wie Geschlecht oder individuelle Lehrerfahrung die empfundene Entlastung und den Leistungsdruck. Es scheinen andere, hier nicht in die Analysen einbezogene Faktoren zum Tragen zu kommen, etwa der Wegfall der zeitintensiven Erstellung der Abituraufgaben oder die Veränderung der Rolle der Lehrpersonen und damit deren Beziehung zu

den Schülerinnen und Schülern (Böhm-Kasper & Weishaupt, 2002; Maag Merki, 2008; van Ackeren et al., 2012; vbw – Vereinigung der Bayerischen Wirtschaft e. V., 2011). Insgesamt zeichnet sich bei den Lehrpersonen kein generelles Muster ab, wer von der Implementation zentraler Abiturprüfungen (nicht) profitiert. Vielmehr lassen sich Unterschiede bei einzelnen Aspekten ausmachen.

Die Annahme differenzieller Effekte der Implementation zentraler Abiturprüfungen auf die *Schülerinnen und Schüler* bestätigt sich hinsichtlich der Abitur- und Halbjahresnoten in Mathematik wie auch mit Blick auf die Emotionen. Von der Einführung des Zentralabiturs in den Mathematik-Leistungskursen profitieren die Schülerinnen und Schüler des ersten Jahrgangs 2008. Sie zeigen signifikant niedrigere Leistungen, erhalten jedoch die höchste Note im Vergleich mit 2007 und 2011. Im ersten Jahr der Reform haben die Lehrpersonen vermutlich als Ausgleich für Unsicherheiten besonders milde benotet. Im längerfristigen Vergleich profitieren im Ausland geborene Schülerinnen und Schüler nicht vom Zentralabitur, da sie trotz gleicher Leistungen eine niedrigere Note im schriftlichen Mathematik-Abitur erhalten. Das unterstützt zwar die Kritik, dass sich zentrale Tests je nach Hintergrund der Schülerinnen und Schüler unterschiedlich auswirken (Amrein & Berliner, 2002b, S. 10ff.; Madaus & Clarke, 2001; Natriello, 2009, S. 1106; Schildkamp et al., 2012; Solórzano, 2008; Thorsen & Cliffordson, 2012). Es darf jedoch nicht übersehen werden, dass der Befund auch für dezentrale Abiturprüfungen im Jahr 2007 Gültigkeit besitzt. Insofern stellt das Zentralabitur weder eine Verbesserung noch eine Verschlechterung dar. Für die Halbjahresnoten ist das Bild weniger einheitlich: In längerfristiger Perspektive (2007-2011) nivellieren sich die Nachteile von Schülern bei einer und die Nachteile von im Ausland geborenen Schülerinnen und Schülern bei zwei Halbjahresnoten. Hingegen profitieren 2011 in zwei Halbjahren Schülerinnen und Schüler mit mehr Büchern zuhause, da sie bei gleicher Leistung eine bessere Note bekommen als diejenigen mit weniger Büchern. Die letztgenannte Gruppe wird somit durch die Einführung des Zentralabiturs benachteiligt.

Da Noten von Lehrpersonen vergeben werden und lediglich bis zu einem gewissen Grad von Schülerinnen und Schülern durch Leistungen und ihr Verhalten beeinflussbar sind, verweisen Noten auf „dahinter" liegende Handlungen der Lehrpersonen sowie auf Mechanismen der (Re-)Produktion von Ungleichheit (Ditton, 2007; Helbig & Nikolai, 2015). Es stellt sich somit auch die Frage, wie Lehrpersonen Noten kommunizieren und wie Schülerinnen und Schüler ihrerseits Noten und Zeugnisse rezipieren (Beutel, 2008, S. 206f.; auch Hattie & Timperley, 2007).

> […] assessment is not an activity that can be done to children, but is accomplished by means of social interaction in which the practices of the participants have a critical effect on the outcome. The outcomes of assessment are actively produced rather than revealed and displayed by the assessment process. (Torrance & Pryor, 2008, S. 236)

Noten und Emotionen stehen in einem komplexen reziproken Verhältnis (Brookhart, 1997; Bürgermeister, 2014). So kann mit Noten etwa das Gefühl von (Un-)Gerechtigkeit oder (mangelnder) Fairness verbunden sein, das je nach persönlichen Merkmalen der Schülerinnen und Schüler variiert (Resh, 2009, 2010). Vor diesem Hintergrund ist von Unterschieden in den längerfristigen Auswirkungen der Einführung zentraler Abiturprüfungen auf das emotionale Erleben in Abhängigkeit vom Leistungsstand der Schülerinnen und Schüler auszugehen. Bei Schülerinnen und Schülern des unteren Quartils steigt die Erfolgsunsicherheit im Abitur von 2007 bis 2011, bei den mittleren Leistungsquartilen bleibt sie stabil und bei den leistungsstarken Schülerinnen und Schülern reduziert sie sich. Dafür empfinden im selben Zeitraum letztere als einzige eine Zunahme der Angst vor Misserfolg. Damit bestätigen sich lediglich zum Teil Forschungsbefunde, die von ungünstigeren (Entwicklungen der) Emotionen leistungsschwächerer Schülerinnen und Schüler ausgehen (Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004). Zudem bestehen Differenzen in der Einschätzung der Unterrichtsgestaltung (Muijs et al., 2005; Rakoczy, 2006; Seifried, 2009; Vanlaar et al., 2016): Die Gruppe der Leistungsstarken erlebt eine Steigerung der Autonomieunterstützung und Motivierungsfähigkeit der Lehrpersonen. Sie scheinen demnach einen stärker unterstützenden und anregenden Unterricht zu erleben als die übrigen Schülerinnen und Schüler (Deci & Ryan, 1993; Rakoczy, 2006; Urhahne, 2015). Diejenigen des mittleren Leistungsniveaus profitieren in längerfristiger Perspektive von einem die Erfolgsunsicherheit im Abitur mildernden Effekt eines positiven Schulklimas. Im Gegenzug löst bei ihnen Autonomieunterstützung stärkere Angst vor Misserfolg aus. Grund dafür könnte ein unausgewogenes Mass an Freiheiten, Anforderungen, Einschätzung der eigenen Fähigkeiten und Attributionen von Erfolg bzw. Misserfolg sein (Boekaerts, 1992; Gläser-Zikuda & Fuß, 2008; Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004; Oerke, 2012b; Oerke et al., 2011; Rakoczy, 2006). Unabhängig vom Leistungsniveau zeigen sich auf deskriptiver Ebene im Sinne der Unterrichtsqualität positive Entwicklungen: Von 2007 zu 2011 nehmen Schülerinnen und Schüler einen Anstieg der Qualität der Vorbereitung auf das Abitur im Unterricht, der Motivierungsfähigkeit der

Lehrpersonen sowie eine grössere Autonomie- und Kompetenzunterstützung wahr. Dies dürfte wiederum die Motivation günstig beeinflussen (Deci & Ryan, 1993). Insgesamt profitieren die Schülerinnen und Schüler vorrangig von einer Erhöhung der Unterrichtsqualität. Die Befunde zu den Noten fallen dagegen uneinheitlich aus.

Die Annahme, dass sich die mit Reformen einhergehenden Veränderungen von Strukturen und Handlungen auf verschiedene Akteursgruppen und Akteure unterschiedlich auswirken (Altrichter & Maag Merki, 2016a, S. 16), findet somit ihre Bestätigung.

7.2 Theoretische Einordnung der empirischen Befunde

Gemäss den Modellen der *Schuleffektivität* wird der Outcome der Schülerinnen und Schüler vom bildungspolitischen Kontext, von Neuausrichtungen von Strukturen und Handlungen auf der Ebene der Einzelschule, von der Unterrichtsqualität sowie von Merkmalen der Schülerinnen und Schüler bestimmt (Creemers & Kyriakides, 2008; Muijs et al., 2005; Muijs et al., 2014; Reynolds, 2005; Reynolds et al., 2015; Scheerens, 1992; Teddlie & Reynolds, 2000). Die Veränderung des bildungspolitischen Kontextes durch die Reform der Abiturprüfungen wirkt sich längerfristig weder auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg der Schülerinnen und Schüler als emotionaler Outcome noch auf den Zusammenhang zwischen Unterrichtdimensionen und Emotionen aus. Auch wenn hier über Veränderungen der Handlungen von Lehrpersonen und deren Effekte auf den Outcome der Schülerinnen und Schüler keine direkten Aussagen möglich sind, lässt sich der Einfluss von kollegialen wie schulischen Merkmalen auf einen Teil der Emotionen der Lehrpersonen in Richtung des Aufbaus einer *school learning environment* oder einer Professionalisierung von Lehrpersonen unter den Bedingungen von *accountability* interpretieren (Hochberg & Desimone, 2010). Dieser Einfluss dürfte sich, vermittelt über den Unterricht, auf den Outcome auswirken. Darauf deutet hin, dass die Schülerinnen und Schüler eine Steigerung der Unterrichtsqualität wahrnehmen.

Anknüpfend an die Unterscheidung von innerer und äusserer Reform (Holtappels, 2014, S. 20f.) lassen sich diese beiden Befunde auch als Indikatoren für eine innere Reform verstehen, während die durch

die Implementation zentraler Abiturprüfungen veränderte Struktur der äusseren Reform zuzurechnen ist. Letztere zählt vor dem Hintergrund der *Educational Governance* zusammen mit Veränderungen von Handlungen zu Effekten erster Ordnung, wohingegen davon ausgehende Transferwirkungen Effekte zweiter Ordnung sind (Altrichter & Maag Merki, 2016a, S. 16). Fast alle in der vorliegenden Arbeit im Fokus stehenden Auswirkungen der Implementation zentraler Abiturprüfungen gehören in die erste Kategorie. Beispielsweise steht das emotionale Erleben der Lehrpersonen ebenso wie das der Schülerinnen und Schüler in Zusammenhang mit modifizierten Strukturen und neuen Beziehungen zwischen Akteursgruppen und zwischen einzelnen Akteuren, welche Neuausrichtungen von Handlungen erfordern. Ist davon die Stabilität der Funktionslogik (Benz, 2009, S. 50) betroffen, kann dies zu Unsicherheiten aufgrund mangelnder Routine, Kontrolle oder Transparenz führen („Repertoire-Unsicherheit"; Rost & Schermer, 1987). Ein Transfereffekt der Anwendung externer, standardisierter Korrekturkriterien bei der Benotung zentraler Abiturprüfungen auf andere Noten, wie die Halbjahresnoten, ließe sich als Effekt zweiter Ordnung einordnen. Zentral ist im Forschungsansatz der Educational Governance die Handlungskoordination zwischen verschiedenen Akteuren und Ebenen (Altrichter, 2015, S. 35). Als Beispiel für die Handlungskoordination verschiedener Akteursgruppen auf der Mikroebene kann gelten, dass die Schülerinnen und Schüler sich von 2007 zu 2011 durch den Unterricht besser auf das Abitur vorbereitet fühlen. In gemeinsamer Abstimmung gelingt es Lehrpersonen sowie Schülerinnen und Schülern, den Unterricht als ko-konstruktiven Prozess (Fend, 2008b) so zu gestalten, dass er Sicherheit in vielfältiger Hinsicht vermittelt – auch wenn dadurch Erfolgsunsicherheit im Abitur und Angst vor Misserfolg der Schülerinnen und Schüler mit der Zeit nicht stärker gemindert werden. Das heisst, dass sich der Effekt nicht unter den Bedingungen dezentraler und zentraler Abiturprüfungen unterscheidet. Dies mag zudem durch das mindestens zweijährige „Arbeitsbündnis" zwischen den beteiligten Akteursgruppen in der Zeit vor dem Abitur begründet sein. Eine andere Erklärung bietet Fend (2008b, S. 21ff.; ähnlich Scheerens, 1992, 14f.), der lediglich von indirekten Auswirkungen des Kontextes auf den ko-konstruktiven Prozess ausgeht.

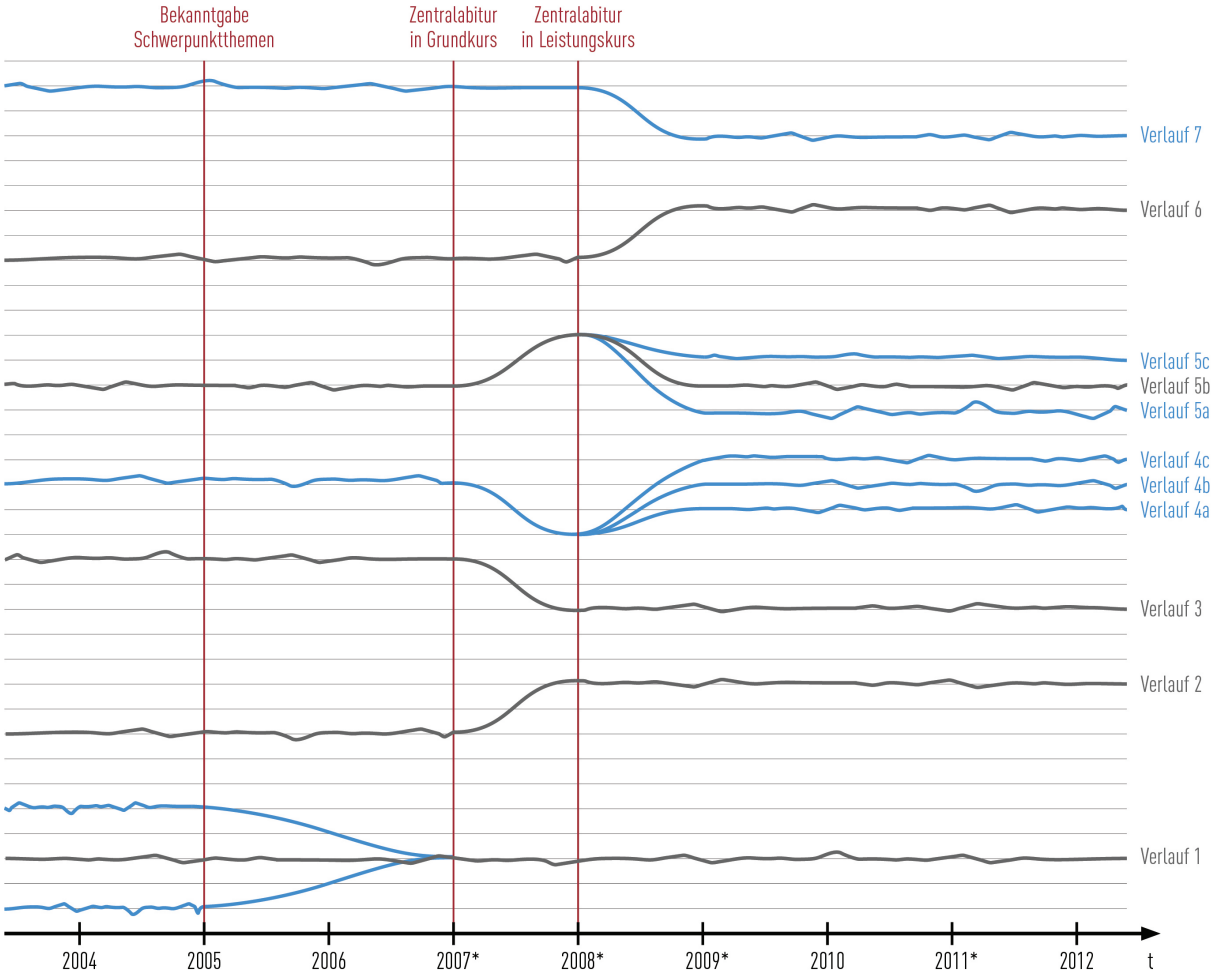
Die Einführung zentraler Abiturprüfungen seitens der Bildungspolitik (Makroebene) kann mit Bezug zur theoretischen und empirischen Auseinandersetzung mit *Schulentwicklung* als Anlass für Veränderungen und Entwicklungen auf der Meso- und Mikroebene gesehen werden (Bischof, 2017, S. 30;

Rolff, 2010, S. 36; van Ackeren et al., 2011, S. 179). Da der Weg dahin womöglich weit ist, empfehlen Berkemeyer et al. (2010, S. 147) Längsschnittstudien zur adäquaten Beobachtung, zur Analyse und zum Verständnis von Schulentwicklungsprozessen. Die hier berichteten Befunde zum emotionalen Erleben der Implementation zentraler Abiturprüfungen einiger Lehrpersonen im Längsschnitt erfüllen diese Empfehlung und tragen zum vertieften Verständnis bei und dies sowohl durch Differenzen als auch durch Bestätigung der Ergebnisse im Querschnitt. Die Differenzen verweisen zudem auf Befunde, nach denen die Kooperation verschiedener Subgruppen von Lehrpersonen eine Grundvoraussetzung für gelingende Schulentwicklung ist (Bennewitz, 2008; Koch, 2009; Oerke, 2012a; Sahner, 2008; Terhart, 2010). Kooperation ist jedoch voraussetzungsreich und bedarf vielfältiger Koordinationsprozesse.

> Durch den tief sitzenden Habitus Unterricht zu machen, werden Blockierungen und Begrenzungen geschaffen, die sich innerhalb der Kooperationsprozesse negativ niederschlagen. Die Fähigkeit sich in angemessener Art und Weise mit Kolleginnen und Kollegen über Aufgaben und Probleme der Steuerung und Entwicklung der Schule und des Unterrichts zu verständigen, ist nicht mit der Fähigkeit gleichzusetzen guten Unterricht zu gestalten. (Bondorf, 2013, S. 193)

Da Lehrpersonen bei der Umsetzung von Reformen eine bedeutende Rolle zukommt, machen auch Personalentwicklungen einen Teil von Schulentwicklung aus (Rolff, 2010, 2013). Unter dem Einfluss vielfältiger Kontextfaktoren wirkt sich die Professionalisierung von Lehrpersonen über Veränderungen von Wissen, Fähigkeiten und beliefs auf den Unterricht und letztlich auf die Leistungen der Schülerinnen und Schüler aus (Hochberg & Desimone, 2010). Die Wahrnehmung der Schülerinnen und Schüler einer längerfristig gesteigerten Vorbereitung auf das Abitur im Unterricht sowie von erhöhter Motivierungsfähigkeit der Lehrpersonen, Autonomie- und Kompetenzunterstützung kann hierfür ein Indikator sein.

Die empirischen Befunde lassen sich unter Rückgriff auf die Weiterentwicklung des Modells der möglichen Entwicklungen/Wirkungen in Abhängigkeit der Zeit (Abb. 4) wie folgt einordnen.



Legende

Modifikation Verlauf Maag Merki (2014)
Erweiterung Elisabeth Maué

*= Jahr der Datenerhebung

Verlauf 1: Stabilität im Zeitverlauf

Verlauf 2: Kurzfristige Zunahme (2007-2008), längerfristige Konsolidierung auf höherem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 3: Kurzfristige Reduktion (2007-2008), längerfristige Konsolidierung auf niedrigerem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 4: Kurzfristige Reduktion (2007-2008), längerfristige Konsolidierung auf
a) niedrigerem, b) höherem, c) gleichem
Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 5: Kurzfristige Zunahme (2007-2008), längerfristige Konsolidierung auf
a) niedrigerem, b) gleichem, c) höherem
Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 6: Kurzfristige Stabilität (2007-2008), verzögerte Zunahme, längerfristige Konsolidierung auf höherem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Verlauf 7: Kurzfristige Stabilität (2007-2008), verzögerte Reduktion, längerfristige Konsolidierung auf niedrigerem Niveau als vor der Einführung zentraler Abiturprüfungen (2007-2011)

Zusätzlich sind bei allen Verläufen Veränderungen (Zunahme oder Reduktion) spätestens mit Bekanntgabe der Schwerpunktthemen im Jahr 2005 für die erste Durchführung zentraler Abiturprüfungen in den Grundkursen 2007 denkbar.

Abbildung 4: Weiterentwicklung des Modells der möglichen Entwicklungen/Wirkungen in Abhängigkeit der Zeit von Maag Merki (2014, S. 63)

Eine Stagnation (*Verlauf 1*) ist zu beobachten beim Zusammenhang der Abitur- und Halbjahresnoten mit dem Leistungstest im Mehrebenenmodell, beim Niveau von Leistungsdruck und Arbeitsunzufriedenheit von Lehrpersonen im Längsschnitt, beim Niveau der Angst vor Misserfolg der Schülerinnen und Schüler und der Erfolgsunsicherheit im Abitur bei denjenigen des mittleren Leistungsniveaus sowie bei fast allen Einflussfaktoren auf die Emotionen der Schülerinnen und Schüler.

Eine kurzfristige Zunahme der Entlastung der Lehrpersonen, die sich anschliessend auf einem höheren Niveau stabilisiert entspricht *Verlauf 2* mit einer Verlängerung des Anstiegs bis zum Jahr 2009.

Spiegelverkehrt reduziert sich, ebenfalls bis zum Jahr 2009, die Unsicherheit der Lehrpersonen und verharrt längerfristig auf einem geringeren Niveau als zu Beginn der Erhebungen (*Verlauf 3*).

Eine kurzfristige Reduktion mit einer längerfristigen Rückkehr zum Ausgangsniveau (*Verlauf 4b*) zeigt sich beim Einfluss des Geburtslandes auf die Abiturnote, bei der Erfolgsunsicherheit im Abitur der Schülerinnen und Schüler sowie beim Effekt der Halbjahresnote 13/1 auf diese Emotion.

Entgegengesetzt dazu verhält sich der kurzfristig auftretende Effekt des familiären Bildungshintergrundes auf die Abiturnote, der sich jedoch nicht längerfristig festigt (*Verlauf 5b*).

Ein verzögert einsetzender Anstieg (*Verlauf 6*) findet sich bei den Korrelationen von Halbjahresnoten und Leistungstest, beim Effekt des familiären Bildungshintergrundes auf zwei Halbjahresnoten, bei der Erfolgsunsicherheit im Abitur der Schülerinnen und Schüler des unteren Leistungsniveaus, der Angst vor Misserfolg derjenigen des oberen Leistungsquartils und bei verschiedenen Aspekten der Unterrichtsqualität.

Verlauf 7 repräsentieren die Reduktion des Effekts des Geschlechts auf eine Halbjahresnote und des Geburtslandes auf zwei Halbjahresnoten, die Verringerung von Leistungsdruck und Arbeitsunzufriedenheit der Lehrpersonen, die Abnahme der Erfolgsunsicherheit im Abitur der leistungsstarken Schülerinnen und Schüler sowie die Nivellierung des Effekts des Schulklimas auf die Erfolgsunsicherheit im Abitur der Schülerinnen und Schüler, allerdings mit einer Ausweitung des Zeitfensters der Reduktion bis zum Jahr 2011.

Die *Verläufe 4a, 4c, 5a und 5c* lassen sich mit den vorliegenden Analysen nicht eindeutig abbilden. Vergleiche der Mittelwerte von Abiturnote und Leistungstest in den Jahren 2007, 2008 und 2011 deuten die Verläufe 4c und 5a an, verfehlen jedoch teilweise das Signifikanzniveau (Publikation 1 im Anhang). Die Einschätzung der Motivierungsfähigkeit der Lehrperson durch Schülerinnen und Schüler des oberen Quartils entspricht Verlauf 5c, wobei die Differenz zwischen den Jahren 2008 und 2011 nicht auf Signifikanz getestet wurde (Publikation 4 im Anhang).

Die gezeigten Verläufe liefern Indizien für Prozesse der *Implementation* der Reform der Prüfungsorganisation seitens der Lehrpersonen. Sind kurzfristig (2007-2008 bzw. 2007-2009) auftretende Effekte in längerfristiger Perspektive (2007-2011) stabil, kann von einer Institutionalisierung der Veränderungen, das heisst einer Integration in die Handlungen der Lehrpersonen, ausgegangen werden (Reynolds, 2005). Beispiele dafür sind die längerfristige Reduktion von Unsicherheit, Leistungsdruck und Arbeitsunzufriedenheit (Verläufe 3 und 7) sowie die längerfristige Zunahme der Entlastung (Verlauf 2). Offenbar konnten die Lehrpersonen neue Handlungsroutinen im Umgang mit den vielfältigen Anforderungen zentraler Abiturprüfungen aufbauen und so Sicherheit gewinnen. Sichtbar werden veränderte Handlungsroutinen in einer längerfristig stärkeren Orientierung der Halbjahresnoten an den Leistungen der Schülerinnen und Schüler (Verlauf 6), in der Reduktion des Effekts des Geschlechts auf eine Halbjahresnote und des Geburtslandes auf zwei Halbjahresnoten. Dieser Standardisierungseffekt des Zentralabiturs auf die Halbjahresnoten kann als Teil von Unterrichtsentwicklung interpretiert werden, ebenso wie die von den Schülerinnen und Schülern eingeschätzte Steigerung der Qualität der Vorbereitung auf das Abitur im Unterricht, der Motivierungsfähigkeit der Lehrpersonen sowie der Autonomie- und Kompetenzunterstützung.

Das Ergebnis ist insofern einzuschränken, als sich in anderen Bereichen abweichende Auswirkungen (z. B. längerfristiger Effekt des familiären Hintergrundes auf zwei Halbjahresnoten) oder keine Veränderungen über die Zeit zeigen (Verlauf 1). Dies steht in Übereinstimmung zu Befunden der Implementation von Bildungsstandards, die darauf hinweisen,

> dass die Frage, wie die Reform ‚Bildungsstandards' in der Schul- und Unterrichtspraxis der Lehrkräfte implementiert wird, nicht von Zustimmung zu der Reform bzw. ihrer Ablehnung auf der Ebene der expliziten Bewertungen und Einstellungen abhängig ist, sondern von der Passung zwischen dem Verständnis, das die Lehrerinnen und Lehrer von den Bildungsstandards entwickeln, und ihren habituellen, implizierten Orientierungen (Asbrand et al., 2012, S. 234).

Reformen werden in Übereinstimmung mit und in Bestätigung von bereits bestehenden Handlungsroutinen gedeutet und legitimiert (Bennewitz, 2008, S. 256f.). Die dadurch ermöglichte Verringerung der Komplexität geht mit der Reproduktion von Handlungen, Strukturen und Einstellungen einher und erschwert dadurch Veränderungen. Für die Reflexion von Handlungen, Strukturen und Einstellungen in Zusammenhang mit der Reform ist die Nähe bzw. Distanz zu persönlichen Erfahrungen sowie zur Professionalität der Lehrpersonen essentiell. Die Bedeutung der Erfahrungen der Lehrpersonen mit zentralen Abiturprüfungen wurde an verschiedenen Stellen der vorliegenden Arbeit gezeigt.

Kurzfristige Veränderungen, die sich längerfristig nicht etablieren, sodass das Ausgangsniveau wie unter den Bedingungen dezentraler Abiturprüfungen wieder erreicht wird, lassen sich mit Bezug zum Konzept des *loose coupling* von Bildungssystemen von Weick (1976) verstehen. In dieses wird zu Beginn einer Reform eingegriffen und das *coupling* durch verstärkte Kontrollen gestrafft, was sich auf das Handeln der Akteure, insbesondere der Lehrpersonen, auswirkt. Mit der Zeit gewinnen die Lehrpersonen jedoch ihre Autonomie oder zumindest einen Teil davon zurück, sodass sich das *coupling* lockert und „alte" Handlungsroutinen wieder an Bedeutung gewinnen (Hamilton et al., 2008; auch Baum, 2014).

Hürden oder „Störquellen" der Implementation stellen mit Kommunikation oder Hierarchien einhergehende Missverständnisse, Widersprüche und Konflikte dar. Zudem möchten nicht unbedingt alle Lehrpersonen etwas verändern, vielmehr versuchen sie, bestehende Strukturen und Handlungen soweit

es geht zu erhalten (Bennewitz, 2008; Brüsemeister, 2010; Holtappels, 2014; Kussau & Brüsemeister, 2007; Spillane et al., 2002; van Ackeren et al., 2011; Zeitler, 2012). Dies könnte eine Erklärung beispielsweise für in weiten Teilen ausbleibende Standardisierungseffekte der Einführung zentraler Abiturprüfungen auf die Benotung sein.

Ein anderer Grund dafür, dass sich Effekte einer *school learning environment* oder einer inneren Reform auf den Outcome nicht durchgängig nachzeichnen lassen, kann darin begründet sein, dass der Zeithorizont von fünf Jahren für die Entstehung kollektiver Lernprozesse der Lehrpersonen, deren Transfer in Handlungen und Handlungsroutinen sowie der Beobachtung der Wirkungen des Handelns zu kurz greift.

Insgesamt verweisen auch die vorliegenden Befunde darauf, dass von einem generellen Effekt der Implementation des Zentralabiturs auf die untersuchten Aspekte nicht auszugehen ist. Vielmehr zeigen sich einzelne, spezifische Effekte, die sich teilweise ergänzen, überlagern, nivellieren oder in Widerspruch zueinander stehen.

Im Fall der Emotionen der Lehrpersonen lassen sich die spezifischen Effekte einer längerfristigen Reduktion von Unsicherheit, Leistungsdruck und Arbeitsunzufriedenheit in Kombination mit der gestiegenen Entlastung als generellen Effekt der Unterstützung der Lehrpersonen in ihrem Handeln interpretieren. Dies ist von bildungspolitischer Seite intendiert (Die Senatorin für Bildung und Wissenschaft, 2013a, S. 4) und als kausaler, starker, tiefgreifender, erwünschter bzw. positiver Effekt zu werten. Dieser generelle Effekt überlagert die nicht-intendierte und unerwünschte Stabilität von Leistungsdruck und Arbeitsunzufriedenheit der Lehrpersonen im Längsschnitt.

Die Steigerung der Unterrichtsqualität in der Wahrnehmung der Schülerinnen und Schüler in längerfristiger Sicht ist als intendierter und erwünschter Effekt zu sehen, auch wenn diese nicht zu einer Minderung von Erfolgsunsicherheit im Abitur und Angst vor Misserfolg führt. Dass sich die Emotionen sowie teilweise auch die Einflüsse auf diese in Abhängigkeit des Leistungsniveaus der Schülerinnen und Schüler unterscheiden, verweist auf einzelne spezifische Effekte. Diese direkten und indirekten Effekte sind nicht intendiert, unerwünscht bzw. negativ und können sich als Effekte zweiter Ordnung in weiteren Bereichen niederschlagen und Unterschiede zwischen Gruppen von Abiturientinnen und Abiturienten vergrössern.

Ebenfalls weder intendiert noch erwünscht sind die kurz- und längerfristig variierenden Einflüsse des Hintergrundes der Schülerinnen und Schüler (Geschlecht, Migrations- und familiärer Hintergrund) sowie die Konstante des Effekts der Leistung auf die Abitur- und Halbjahresnoten. Einzelne, spezifische, intendierte und erwünschte bzw. positive Effekte stehen anderen spezifischen, nichtintendierten und unerwünschten bzw. negativen Effekten gegenüber und überlagern diese zum Teil. Die längerfristig verengte Korrelation zwischen Leistungstest und Halbjahresnoten gilt als schwacher, spezifischer, indirekter, erwünschter, aber nicht unbedingt erwarteter Effekt (Altrichter & Maag Merki, 2016a, S. 16; Baum, 2014, S. 248; Koch, 2009, S. 134f.; Reynolds, 2005, S. 13; Rogers, 2003, S. 30f.; Watanabe, 2004, S. 20ff.).

Somit spiegeln die einzelnen Befunde die im theoretischen Hintergrund (Kapitel 3) dargelegte Komplexität der Einflüsse und Auswirkungen auf das über mehrere Ebenen, Akteursgruppen und individuelle Akteure interrelational verwobene System von Schule und Unterricht wider und offenbaren die Begrenzung der Steuerungsmöglichkeiten. Darüber hinaus verstärken sie das Fazit des empirischen Hintergrundes (Kapitel 4), dass die Effekte der Reform der Prüfungsorganisation am Ende der Sekundarstufe II in Deutschland je nach Bundesland, Schule, Kursniveau und Fach – sowie Akteursgruppe und Akteure – differieren.

7.3 Limitationen

Die Analyse längerfristiger Auswirkungen der Implementation zentraler Abiturprüfungen in Bremen sowohl auf die Vergleichbarkeit der Abitur- und Halbjahresnoten als auch auf das emotionale Erleben von Lehrpersonen, Schülerinnen und Schülern erfolgt nicht ohne Limitationen. Neben den Limitationen der einzelnen Analysen, auf die in den jeweiligen Beiträgen (Kapitel 6; Publikationen im Anhang) eingegangen wird, gibt es weitere Limitationen.

Die Datenstruktur bedingt die Verwendung ausgewählter Stichproben in den Analysen. Lediglich für die Leistungskurse liegen Daten von dezentralen (2007) und zentralen Abiturprüfungen vor (ab 2008), welche den Vergleich der beiden Prüfungsformen ermöglichen. Insofern stützen sich die Analysen der Emotionen der Schülerinnen und Schüler ebenso wie die der Noten auf Leistungskurse. Bei den Noten

kommt hinzu, dass ausschliesslich das Fach Mathematik betrachtet werden kann, da im Jahr 2011 einzig in Mathematik ein externer, standardisierter Leistungstest durchgeführt wurde.

Die Analysen des emotionalen Erlebens der Implementation zentraler Abiturprüfungen der Akteursgruppen Lehrpersonen wie Schülerinnen und Schüler konzentrieren sich vorrangig auf „negative" Emotionen und vernachlässigen „positive" Emotionen wie Freude. Letztere müssten in weiteren Analysen in den Blick genommen werden. Die Emotionen von Lehrpersonen wurden teils lediglich mittels inhaltlich breit formulierter Einzelitems erhoben, aus denen nicht direkt hervorgeht, weshalb Lehrpersonen beispielsweise Leistungsdruck oder Entlastung empfinden.

Neben der Untersuchung des emotionalen Erlebens der Lehrperson stehen deren Handlungen sowohl bei der Vergabe der Noten als auch bei den Auswirkungen des Unterrichts auf die Emotionen der Schülerinnen und Schüler zumindest indirekt im Fokus. Die Befunde geben zwar Hinweise auf potentielle Handlungen, ohne dass jedoch Aussagen über tatsächlich ablaufende komplexe Prozesse der Interpretation, Adaption, Rekontextualisierung und Bewältigung der Reform der Prüfungsorganisation in Abhängigkeit von Erfahrungswissen und Handlungsroutinen der Lehrpersonen oder schulspezifischen Merkmalen wie der Schulkultur sowie von Koordinationsprozessen zwischen einzelnen Akteursgruppen und Ebenen möglich sind (Bormann, 2011; Rustemeyer, 2009; Zeitler, 2012). Für die Analyse innerer Verarbeitungsprozesse äusserer Anregungen durch Veränderung der Rahmenbedingungen böten sich tiefergehende qualitative Erhebungen als Ergänzung der quantitativen Erhebungen an.

7.4 Implikationen

Unter Berücksichtigung der Limitationen lassen sich für die Erforschung von Reformvorhaben ebenso wie für deren praktische Umsetzung Implikationen ableiten.

Implikationen für die Forschung zu Reformen

Die Vielzahl und Vielfalt spezifischer Effekte der Implementation zentraler Abiturprüfungen in Bremen, die oftmals in kurzfristiger (2007-2008 bzw. 2007-2009) und längerfristiger Perspektive (2007-2011) variieren, verweisen auf die Bedeutung eines geeigneten Zeithorizonts zur Nachzeichnung von

Auswirkungen einer Reform. Um Rahmenbedingungen und Handlungen verschiedener Akteure und Ebenen zum Zeitpunkt vor und nach der Reform in Relation zueinander setzen zu können, ist eine adäquate Berücksichtigung der Zeit vor der Reform notwendig (Abbildung 3). Nicht erst mit der erstmaligen Durchführung einer initiierten Änderung wirkt sich eine Reform aus, sondern vermutlich bereits in ihrer bildungspolitischen, wissenschaftlichen wie gesellschaftlichen Diskussion, spätestens jedoch mit der endgültigen Entscheidung dafür.

Die verschiedenen Ebenen des Bildungssystems stehen in komplexer Interrelation und Interaktion. Insofern müssen zur umfassenden Erforschung von Effekten einer Reform die direkt betroffenen Ebene(n) mit verschiedenen Akteuren wie auch die potentiell indirekt betroffenen Ebene(n) und Akteure einbezogen werden. Sie lassen sich anhand der Intentionen einer Reform und der Strukturen und Handlungen bestimmen, die für deren Umsetzung erforderlichen sind. Im Bildungsbereich zielen Reformen häufig auf die Verbesserung der Lernleistungen der Schülerinnen und Schüler ab und erfordern damit vorrangig von den Lehrpersonen auf der Mikro- und Mesoebene den Aufbau neuer Strukturen sowie Neuausrichtungen der Handlungen. Lehrpersonen sind als einzelne Akteure wie auch als Kollegium in Reformen eingebunden.

Darüber hinaus gilt es, verschiedene Aspekte, bei denen direkte und indirekte Auswirkungen einer Reform zu vermuten sind, in der nötigen inhaltlichen Breite und Tiefe abzudecken. Erst dann ergibt sich ein Gesamtbild verschieden intensiver, spezifischer, nicht-intendierter, (in)direkter, (un)erwarteter Effekte einer Reform, die sich überlagern, ergänzen, nivellieren oder in Widerspruch zueinander stehen können (Altrichter & Maag Merki, 2016a; Baum, 2014; Koch, 2009; Reynolds, 2005; Rogers, 2003; Watanabe, 2004). Dafür bietet es sich an, ein triangulatives Forschungsdesign aus einer Kombination quantitativer und qualitativer Erhebungen zu wählen und so die Vorteile der jeweiligen Methoden auszuschöpfen.

Implikationen für die Praxis

Die Implementation einer Reform im Sinne der Transformation in Handlungsroutinen erfordert sowohl eine Übereinstimmung mit den Zielen als auch Änderungen von Einstellungen, Werten, beliefs oder des Habitus von Akteuren (Hopkins et al., 1994; Reynolds, 2005). In einem ersten Schritt müssen die Akteure die Ziele und Neuerungen der Reform verstehen und in einem zweiten Schritt Einstellung, Verhalten und

Nutzen akzeptieren (Penninckx et al., 2017; Spillane et al., 2002; Zeitler, 2012; Ziegelbauer, 2015). Hierfür ist eine intensive Kommunikation von grosser Bedeutung. Neben Ressourcen bedarf es vor allem einer Begleitung der Akteure, etwa durch Massnahmen der Professionalisierung. Diese sollten in Kongruenz zur Reform stehen sowie an die individuellen und kollektiven Rahmenbedingungen auf der Mikro- und Mesoebene anschliessen. Daraus lässt sich ableiten, dass identische Professionalisierungsmassnahmen nicht für alle Akteure passend sind. Mit Blick auf die Implementation der Bildungsstandards zeigen Asbrand et al. (2012), dass Lehrpersonen je nach Erfahrung mit Unterrichtsentwicklung und Habitus unterschiedliche Strategien benötigen bzw. anwenden.

In diese Richtung weist auch der Befund der vorliegenden Arbeit, dass nicht alle Akteure in gleichem Mass von der Implementation zentraler Abiturprüfungen profitieren. Dies könnte in einem Ungleichgewicht der Anforderungen einer Reform und der persönlichen Ressourcen für deren Bewältigung begründet sein. Diese fehlende Balance sollte hergestellt werden durch ein adäquates Verständnis der Intentionen der Reform oder durch die Reflexion der eigenen Handlungsroutinen und es sollten Anstösse zu deren Modifikation gegeben werden, ausgerichtet an den Bedürfnissen der jeweiligen Akteure. So wird der Wechsel vom „disequilibrium" zum „dynamic equilibrium" (Rogers, 2003, S. 471) ermöglicht, was einer Passung der Veränderungen mit den Ressourcen der Akteure entspricht. Diese wird sich in modifizierten Handlungen niederschlagen, im vorliegenden Fall etwa in einer verstärkten Orientierung der Notengebung an externen Kriterien oder in einer Erhöhung der Unterrichtsqualität.

Auf die Bedeutung schulischer und kollegialer Aspekte für die Bewältigung der mit der Implementation zentraler Abiturprüfungen einhergehenden Unsicherheiten verweisen Befunde der vorliegenden Arbeit (Kapitel 6.3; Publikation 3 im Anhang). Hier könnten Massnahmen der Organisations-, Schul-, Unterrichtsentwicklung ansetzen und zur Stärkung des sozialen Kapitals von Lehrpersonen beitragen (Bondorf, 2013; Maag Merki, 2016; Penuel et al., 2012; Rolff, 2010).

Reformen wirken sich jedoch nicht ausschliesslich unidirektional auf verschiedene Ebenen sowie auf Akteure und Akteurskonstellationen aus, sondern sie werden auch ihrerseits verändert (Rogers, 2003; Tyack & Tobin, 1994). Das verweist zum einen auf sense-making-Prozesse, zum anderen auf die Möglichkeit, Modifikationen einer Reform bottom-up zu initiieren (Spillane et al., 2002). Aufgrund

des relativ starken Einbezugs der Lehrpersonen in die Belange zentraler Abiturprüfungen sowie des speziellen Monitoring- und Feedbackprozesses (Kapitel 2.2) können Lehrpersonen ihre Erfahrungen mit und Sichtweisen auf das Zentralabitur rückmelden und Vorschläge für Änderungen oder Wünsche nach Konstanten einbringen. Diesbezüglich begünstigen die geringe Grösse und die niedrige Anzahl an Schulen mit gymnasialer Oberstufe des Bundeslandes Bremen die Möglichkeiten der Kommunikation und Interaktion untereinander auf Ebene der Lehrpersonen, der Einzelschulen wie auch über unterschiedliche Ebenen hinweg.

8. Fazit

Gegenstand der vorliegenden Arbeit ist die Analyse von Auswirkungen der Implementation zentraler Abiturprüfungen im Bundesland Bremen auf die Akteursgruppen Lehrpersonen sowie Schülerinnen und Schüler über einen Zeitraum von fünf Jahren. Im Fokus stehen die Vergleichbarkeit der Abitur- und Halbjahresnoten in Mathematik-Leistungskursen, das emotionale Erleben des Zentralabiturs der beiden Akteursgruppen und die Frage, ob Personen von der Reform der Prüfungsorganisation profitieren. Die inhaltlichen Dimensionen verweisen auf mit der Reform einhergehende Effekte, die seitens der Bildungspolitik intendiert sind – etwa die Steigerung der Standardisierung und Vergleichbarkeit der Noten – wie nicht-intendiert sind – beispielsweise eine Erhöhung von Stress und Druck für die Beteiligten.

In Übereinstimmung zu bisherigen Forschungsbefunden lässt sich auch in längerfristiger Perspektive (2007 bis 2011) kein genereller Effekt der Implementation des Zentralabiturs ausmachen. Vielmehr zeigen sich einzelne, spezifische Effekte, die sich beim emotionalen Erleben der Lehrpersonen zu einem generellen Effekt entlastender Unterstützung verdichten. Davon abgesehen stehen positive und negative Effekte sowie Konstanten nebeneinander und überlagern sich zum Teil.

Zentrale Abiturprüfungen finden am Ende der schulischen Laufbahn von Schülerinnen und Schülern statt und weisen damit lediglich einen begrenzten Anteil an der gesamten schulischen „Lernkarriere" auf. Dennoch spielen sie aufgrund der Bedeutung des Abiturs für den weiteren Lebensweg eine entscheidende Rolle und sollten den von unterschiedlichen Akteuren der Makro-, Meso- und Mikroebene gestellten Ansprüchen genügen. Die schrittweise Implementation des Zentralabiturs im Bundesland Bremen birgt viel Potenzial, was jedoch noch nicht ausreichend ausgeschöpft ist.

Literaturverzeichnis

Abs, H. J., Brüsemeister, T., Schemmann, M., & Wissinger, J. (2015). Akzentsetzungen bei der Erforschung von Steuerung und Koordination in Mehrebenensystemen. In H. J. Abs, T. Brüsemeister, M. Schemmann, & J. Wissinger (Hrsg.), *Governance im Bildungssystem. Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 7-17). Wiesbaden: Springer VS.

Allen, A. (2012). Cultivating the myopic learner: the shared project of high-stakes and low-stakes assessment. *British Journal of Sociology of Education, 33*(5), 641-659. doi:10.1080/01425692.2012.668832

Altrichter, H. (2015). Governance – Steuerung und Handlungskoordination bei der Transformation von Bildungssystemen. In H. J. Abs, T. Brüsemeister, M. Schemmann, & J. Wissinger (Hrsg.), *Governance im Bildungssystem. Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 21-63). Wiesbaden: Springer VS.

Altrichter, H., Brüsemeister, T., & Wissinger, J. (2007). Einführung. In H. Altrichter, T. Brüsemeister, & J. Wissinger (Hrsg.), *Educational Governance. Handlungskoordination und Steuerung im Bildungssystem* (S. 9-13). Wiesbaden: VS.

Altrichter, H., & Maag Merki, K. (2016a). Steuerung der Entwicklung des Schulwesens. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (2., überarbeitete und aktualisierte Auflage, S. 1-27). Wiesbaden: Springer VS.

Altrichter, H., & Maag Merki, K. (Hrsg.). (2010). *Handbuch neue Steuerung im Schulsystem*. Wiesbaden: VS.

Altrichter, H., & Maag Merki, K. (Hrsg.). (2016b). *Handbuch Neue Steuerung im Schulsystem* (2., überarbeitete und aktualisierte Auflage). Wiesbaden: Springer VS.

Altrichter, H., Moosbrugger, R., & Zuber, J. (2016a). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (2., überarbeitete und aktualisierte Auflage, S. 235-277). Wiesbaden: Springer VS.

Altrichter, H., Rürup, M., & Schuchart, C. (2016b). Schulautonomie und die Folgen. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulwesen* (2., überarbeitete und aktualisierte Auflage, S. 107-149). Wiesbaden: Springer VS.

Amengual Pizarro, M. (2010). Exploring the Washback Effects of a High-Stakes English Test on the Teaching of English in Spanish Upper Secondary Schools. *Revista Alicantina de Estudios Ingleses, 23*, 149-170. doi:10.14198/raei.2010.23.09

Amrein, A. L., & Berliner, D. C. (2002a). *An analysis of some unintended and negative consequences of high-stakes testing*. Tempe, AZ: Education Policy Research Unit (EPRU), Education Policy Studies Laboratory, College of Education, Arizona State University.

Amrein, A. L., & Berliner, D. C. (2002b). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). doi:10.14507/epaa.v10n18.2002

Appius, S. (2012). Kooperationen zwischen Lehrpersonen im Zusammenhang mit dem Abitur. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 91-113). Wiesbaden: Springer VS.

Appius, S., & Holmeier, M. (2012). Beurteilung der Abituraufgaben und Korrekturhinweise. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 353-381). Wiesbaden: Springer VS.

Argyris, C., & Schön, D. A. (1999). *Die Lernende Organisation. Grundlagen, Methode, Praxis*. Stuttgart: Klett-Cotta.

Asbrand, B., Zeitler, S., & Heller, N. (2012). Diskussion und Ausblick. In S. Zeitler, N. Heller, & B. Asbrand (Hrsg.), *Bildungsstandards in der Schule. Eine rekonstruktive Studie zur Implementation der Bildungsstandards* (S. 231-254). Münster et al.: Waxmann.

Bandelow, N. C. (2004). Governance im Gesundheitswesen: Systemintegration zwischen Verhandlung und hierarchischer Steuerung. In S. Lange, & U. Schimank (Hrsg.), *Governance und gesellschaftliche Integration* (S. 89-107). Wiesbaden: VS.

Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: The engine of systemic curricular reform? *Journal of curriculum studies, 32*(5), 623-650. doi:10.1080/00220270050116923

Bastian, J. (2007). *Einführung in die Unterrichtsentwicklung*. Weinheim & Basel: Beltz.

Baum, E. (2014). *Kooperation und Schulentwicklung. Wie Lehrkräfte in Gruppen Entwicklungsanlässe bearbeiten*. Wiesbaden: Springer VS. doi:10.1007/978-3-531-19025-9

Baumert, J. (2016). Leistungen, Leistungsfähigkeit und Leistungsgrenzen der empirischen Bildungsforschung: Das Beispiel von Large-Scale-Assessment-Studien zwischen Wissenschaft und Politik. *Zeitschrift für Erziehungswissenschaft, 19*(Supplement 1), 215-253. doi:10.1007/s11618-016-0704-4

Baumert, J., & Watermann, R. (2000). Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. In J. Baumert, W. Bos, & R. Lehmann (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 317-372). Opladen: Leske + Budrich.

Becker, E. S., Goetz, T., Morger, V., & Ranellucci, J. (2014). The importance of teachers' emotions and instructional behavior for their students' emotions – An experience sampling analysis. *Teaching and Teacher Education, 43*, 15-26. doi:10.1016/j.tate.2014.05.002

Becker, E. S., Keller, M. M., Goetz, T., Frenzel, A. C., & Taxer, J. L. (2015). Antecedents of teachers' emotions in the classroom: an intraindividual approach. *Frontiers in Psychology, 6*(Article 635), 1-12. doi:10.3389/fpsyg.2015.00635

Bellmann, J. (2016). Output- und Wettbewerbssteuerung im Schulsystem. Konzeptionelle Grundlagen und empirische Befunde. In M. Heinrich, & B. Kohlstock (Hrsg.), *Ambivalenzen des Ökonomischen. Analysen zur „Neuen Steuerung" im Bildungssystem* (S. 13-34). Wiesbaden: Springer VS.

Bennewitz, H. (2008). Lehrende in Schulreformprozessen. Eine Deutungsmusteranalyse. In G. Breidenstein, & F. Schütze (Hrsg.), *Paradoxien in der Reform der Schule. Ergebnisse qualitativer Sozialforschung* (S. 247-260). Wiesbaden: VS.

Benz, A. (2009). *Politik in Mehrebenensystemen*. Wiesbaden: VS.

Benz, A., Lütz, S., Schimank, U., & Simonis, G. (2007). Einleitung. In A. Benz, S. Lütz, U. Schimank, & G. Simonis (Hrsg.), *Handbuch Governance. Theoretische Grundlagen und empirische Anwendungsfelder* (S. 9-25). Wiesbaden: VS.

Berkemeyer, N., Bos, W., & Gröhlich, C. (2010). Schulentwicklungsprozesse in Längsschnittstudien. In T. Bohl, W. Helsper, H. G. Holtappels, & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie – Forschungsbefunde – Entwicklungsprozesse – Methodenrepertoire* (S. 147-150). Bad Heilbrunn: Julius Klinkhardt.

Berner, E., Oelkers, J., & Reusser, K. (2008). Implementation von Bildungsstandards: Bedingungen des Gelingens (und Scheiterns) aus internationaler Sicht. *Zeitschrift für Pädagogik, 53. Beiheft*, 210-226.

Berry, R., & Adamson, B. (2011). Assessment Reform Past, Present and Future. In R. Berry, & B. Adamson (Hrsg.), *Assessment Reform in Education. Policy and Practice* (S. 3-14). Dordrecht et al: Springer.

Beutel, S.-I. (2008). Das Wissen der Kinder über die Schule und ihr Lernen einbeziehen – Neuere Forschungen zur Leistungsbeurteilung. In G. Breidenstein, & F. Schütze (Hrsg.), *Paradoxien in der Reform der Schule. Ergebnisse qualitativer Sozialforschung* (S. 201-215). Wiesbaden: VS.

Bieri, T. (2006). *Lehrpersonen: Hoch belastet und trotzdem zufrieden?* Bern et al.: Haupt.

Biermann, F., & Pattberg, P. (2004). Governance zur Bewahrung von Gemeinschaftsgütern. Grundprobleme und Institutionen der Umweltpolitik. In S. Lange, & U. Schimank (Hrsg.), *Governance und gesellschaftliche Integration* (S. 169-187). Wiesbaden: VS.

Bischof, L. M. (2017). *Schulentwicklung und Schuleffektivität. Ihre theoretische und empirische Verknüpfung*. Wiesbaden: Springer VS.

Bishop, J. H. (1995). The impact of curriculum-based external examinations on school priorities and student learning. *International Journal of Educational Research, 23*(8), 653-752. doi:10.1016/0883-0355(96)00001-8

Bishop, J. H. (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review, 6*, 349-398.

Bishop, J. H., & Wößmann, L. (2004). Institutional Effects in a Simple Model of Educational Production. *Education Economics, 12*(1), 17-38. doi:10.1080/0964529042000193934

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451-469. doi:10.1080/0969594x.2011.557020

Blättel-Mink, B., & Menez, R. (2015). *Kompendium der Innovationsforschung* (2. Auflage). Wiesbaden: Springer VS.

Block, R., Klein, E. D., van Ackeren, I., & Kühn, S. M. (2011). Leistungseffekte des Zentralabiturs? Eine kritische Auseinandersetzung mit bildungsökonomischen Interpretationen zu den Effekten der Prüfungsorganisation auf der Basis von PISA-E-2003-Daten. *Bildungsforschung, 8*(1), 215-238.

Bode, I. (2010). Disorganisierte Governance und Unterpriviligierung. Die Konsequenzen neuer Steuerungsformen in der gesetzlichen Krankenkasse. In U. Clement, J. Nowak, C. Scherrer, & S. Ruß (Hrsg.), *Public Governance und schwache Interessen* (S. 27-46). Wiesbaden: VS.

Boekaerts, M. (1992). The Adaptable Learning Process: Initiating and Maintaining Behavioural Change. *Applied Psychology: An International Review, 41*(4), 377-397.

Böhm-Kasper, O. (2004). *Schulische Belastung und Beanspruchung. Eine Untersuchung von Schülern und Lehrern am Gymnasium*. Münster et al.: Waxmann.

Böhm-Kasper, O., & Weishaupt, H. (2002). Belastung und Beanspruchung von Lehrern und Schülern am Gymnasium. *Zeitschrift für Erziehungswissenschaft, 5*(3), 472-499.

Bondorf, N. (2013). *Profession und Kooperation eine Verhältnisbestimmung am Beispiel der Lehrerkooperation*. Wiesbaden: Springer VS.

Bonsen, M. (2005). Professionelle Lerngemeinschaften in der Schule. In H. G. Holtappels, & K. Höhmann (Hrsg.), *Schulentwicklung und Schulwirksamkeit. Systemsteuerung, Bildungschancen und Entwicklung der Schule* (S. 180-195). Weinheim & München: Juventa.

Bormann, I. (2011). Innovationen als ‚Wissenspassagen'. Theoretische Grundlegung und Implikationen für die Analyse. *Die deutsche Schule, 103*(1), 53-64.

Bormann, I. (2012). Vertrauen in Institutionen der Bildung oder: Vertrauen ist gut – ist Evidenz besser? *Zeitschrift für Pädagogik, 58*(6), 812-823.

Bormann, I., & Hamborg, S. (2015). Transfer und Institutionalisierung im Bildungsbereich. Einblicke in eine governance-analytische Triangulationsstudie. In H. J. Abs, T. Brüsemeister, M. Schemmann, & J. Wissinger (Hrsg.), *Governance im Bildungssystem. Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 291-314). Wiesbaden: Springer VS.

Bornkessel, P., & Kuhnen, S. U. (2011). Zum Einfluss der sozialen Herkunft auf Schulleistung, Studienzuversicht und Studienintention am Ende der Sekundarstufe II. In P. Bornkessel, & J. Asdonk (Hrsg.), *Der Übergang Schule – Hochschule. Zur Bedeutung sozialer, persönlicher und institutioneller Faktoren am Ende der Sekundarstufe II* (S. 47-104). Wiesbaden: VS.

Böttcher, W. (2005). Chancengleichheit als Herausforderung. Oder: Wie an einem Problem vorbei agiert wird. In H. G. Holtappels, & K. Höhmann (Hrsg.), *Schulentwicklung und Schulwirksamkeit. Systemsteuerung, Bildungschancen und Entwicklung der Schule* (S. 99-109). Weinheim & München: Juventa.

Böttcher, W. (2012). Zur Kritik des Regierens in der Schulpolitik. Zentralisierung und Vertrauen statt Dezentralisierung und Kontrolle. In S. Hornberg, & M. Parreira do Amaral (Hrsg.), *Deregulierung im Bildungswesen* (S. 29-52). Münster et al.: Waxmann.

Böttcher, W., Dicke, J. N., & Ziegler, H. (2012). Erziehungswissenschaft, Bildungspolitik und Bildungspraxis. Anmerkungen zu einem schwierigen Verhältnis. In W. Böttcher, J. N. Dicke, & H. Ziegler (Hrsg.), *Evidenzbasierte Bildung. Wirkungsevaluation in Bildungspolitik und pädagogischer Praxis* (S. 7-21). Münster et al.: Waxmann.

Bourdieu, P. (1982). *Die feinen Unterschiede. Kritik der gesellschaftlichen Urteilskraft*. Frankfurt am Main: Suhrkamp.

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation, 17*(3), 141-159. doi:10.1080/13803611.2011.597112

Braun, D. (2004). Wie nützlich darf Wissenschaft sein? Zur Systemintegration von Wissenschaft, Ökonomie und Politik. In S. Lange, & U. Schimank (Hrsg.), *Governance und gesellschaftliche Integration* (S. 65-87). Wiesbaden: VS.

Brimijoin, K. (2005). Differentiation and High-Stakes Testing: An Oxymoron? *Theory Into Practice, 44*(3), 254-261. doi:10.1207/s15430421tip4403_10

Brookhart, S. M. (1997). A Theoretical Framework for the Role of Classroom Assessment in Motivating Student Effort and Achievement. *Applied Measurement in Education, 10*(2), 161-180. doi:10.1207/s15324818ame1002_4

Brüsemeister, T. (2010). Educational Governance – Aufriss von Perspektiven für die Empirische Bildungsforschung. In C. Hof, J. Ludwig, & B. Schäffer (Hrsg.), *Steuerung – Regulation – Gestaltung. Governance-Prozesse in der Erwachsenenbildung zwischen Struktur und Handlung* (S. 7-16). Hohengehren: Schneider.

Brüsemeister, T., Altrichter, H., & Heinrich, M. (2010). Governance und Schulentwicklung. In T. Bohl, W. Helsper, H. G. Holtappels, & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie – Forschungsbefunde – Entwicklungsprozesse – Methodenrepertoire* (S. 126-129). Bad Heilbrunn: Julius Klinkhardt.

Bryk, A. S. (2010). Organizing Schools for Improvement. *Phi Delta Kappan, 91*(7), 23-30. doi:10.2307/25655236

Büchel, F., Jürges, H., & Schneider, K. (2003). Die Auswirkungen zentraler Abschlussprüfungen auf die Schulleistung – Quasi-experimentelle Befunde aus der deutschen TIMSS-Stichprobe. *Vierteljahreshefte zur Wirtschaftsforschung, 72*(2), 238-251.

Buchwald, P. (2011). *Stress in der Schule und wie wir ihn bewältigen*. Paderborn: Schöningh.

Bürgermeister, A. (2014). *Leistungsbeurteilung im Mathematikunterricht. Bedingungen und Effekte von Beurteilungspraxis und Beurteilungsgenauigkeit*. Münster & New York: Waxmann.

Buske, R. (2014). *Die Entwicklung kollektiver Innovationsbereitschaft von Lehrkollegien. Eine theoretische Modellierung und empirische Untersuchung im Längsschnitt*. Landau: Verlag Empirische Pädagogik.

Camilli, G., & Monfils, L. F. (2004). Test Scores and Equity. In W. A. Firestone, R. Y. Schorr, & L. F. Monfils (Hrsg.), *The Ambiguity of Teaching to the Test. Standards, Assessment, and Educational Reform* (S. 143-157). Mahwah, NJ: Lawrence Erlbaum Associates.

Cappellari, L., Lucifora, C., & Pozzoli, D. (2012). Determinants of grades in maths for students in economics. *Education Economics, 20*(1), 1-17. doi:10.1080/09645291003718340

Cheng, L., & Sun, Y. (2015). Teachers' Grading Decision Making: Multiple Influencing Factors and Methods. *Language Assessment Quarterly, 12*, 213-233. doi:10.1080/15434303.2015.1010726

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: research contexts and methods*. Mahwah, NJ: Wawrence Erlbaum Associates.

Christ, O., & Schlüter, E. (2012). *Strukturgleichungsmodelle mit Mplus. Eine praktische Einführung*. München: Oldenburg.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NY: Erlbaum.

Collie, R. J., Shapka, J. D., & Perry, N. E. (2012). School Climate and Social-Emotional Learning: Predicting Teacher Stress, Job Satisfaction, and Teaching Efficacy. *Journal of Educational Psychology, 104*(4), 1189-1204.

Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.

Creemers, B. P. M. (o. J.). *The Comprehensive Model of Educational Effectiveness. Background, major assumptions and description*. Retrieved from https://www.rug.nl/staff/b.p.m.creemers/the_comprehensive_model_of_educational_effectiveness.pdf

Creemers, B. P. M., & Kyriakides, L. (2008). *The Dynamics of Educational Effectiveness. A contribution to policy, practice and theory in contemporary schools*. London & New York: Routledge.

Creemers, B. P. M., & Kyriakides, L. (2010). School Factors Explaining Achievement on Cognitive and Affective Outcomes: Establishing a Dynamic Model of Educational Effectiveness. *Scandinavian Journal of Educational Research, 54*(3), 263-294. doi:10.1080/00313831003764529

Creemers, B. P. M., Kyriakides, L., & Antoniou, P. (2013a). A dynamic approach to school improvement: main features and impact. *School leadership & management, 33*(2), 114-132. doi:10.1080/13632434.2013.773883

Creemers, B. P. M., Kyriakides, L., Panayiotou, A., Bos, W., Holtappels, H. G., Pfeifer, M., Vennemann, M., Wendt, H., Scharenberg, K., Smyth, E., McMahon, L., McCoy, S., Van Damme, J., Vanlaar, G., Antoniou, P., Charalambous, C., Charalambous, E., Maltezou, E., Zupanc, D., Bren, M., Cankar, G., Hauptman, A., Rekalidou, G., Penderi, E., Karadimitriou, K., Dimitriou, A., Desli, D., & Tempridou, A. (2013b). *Establishing a knowledge base for quality in education: Testing a dynamic theory for education. Handbook on designing evidence-based strategies and actions to promote quality in education*. Münster et al.: Waxmann.

Dalin, P. (1999). *Theorie und Praxis der Schulentwicklung*. Neuwied, Kriftel: Luchterhand.

Day, C., Elliot, B., & Kington, A. (2005). Reform, standards and teacher identity: Challenges of sustaining commitment. *Teaching and Teacher Education, 21*, 563-577. doi:10.1016/j.tate.2005.03.001

Deci, E. L., & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik, 39*(2), 223-238.

Deutscher Bundestag. (Hrsg.). (2016). *Grundgesetz für die Bundesrepublik Deutschland*. Retrieved from https://www.btg-bestellservice.de/pdf/10060000.pdf

Die Senatorin für Bildung und Wissenschaft. (2010). *Richtlinie für die Aufgabenstellung und Bewertung der Leistungen in der Abiturprüfung (ARI) vom 1. Februar 2008 in der Fassung vom 15. Oktober 2010*. Retrieved from https://www.bildung.bremen.de/sixcms/media.php/13/ARI.pdf

Die Senatorin für Bildung und Wissenschaft. (2013a). *Abiturprüfung 2015. Regelungen für das erste bis dritte Prüfungsfach mit landesweit einheitlicher Aufgabenstellung*. Retrieved from www.lis.bremen.de/sixcms/media.php/13/v_10-2013_a.55413.pdf

Die Senatorin für Bildung und Wissenschaft. (2013b). *Lesefassung der Verordnung über die Abiturprüfung im Lande Bremen (AP-V) vom 01.12.05 in der Fassung vom 1. August 2007*. Retrieved from https://www.bildung.bremen.de/sixcms/media.php/13/Lesefassung2013.pdf

Diemer, T., & Kuper, H. (2011). Formen innerschulischer Steuerung mittels zentraler Lernstandserhebungen. *Zeitschrift für Pädagogik, 57*(4), 554-571.

Ditton, H. (2000). Qualitätskontrolle und Qualitätssicherung in Schule und Unterricht. Ein Überblick zum Stand der empirischen Forschung. In A. Helmke, W. Hornstein, & E. Terhart (Hrsg.), *Qualität und Qualitätssicherung im Bildungsbereich: Schule, Sozialpädagogik, Hochschule* (S. 73-92). Weinheim & Basel: Beltz.

Ditton, H. (2007). Der Beitrag von Schule und Lehrern zur Reproduktion von Bildungsungleichheit. In R. Becker, & W. Lauterbach (Hrsg.), *Bildung als Privileg. Erklärungen und Befunde zu den Ursachen der Bildungsungleichheit* (2., aktualisierte Auflage, S. 243-271). Wiesbaden: VS.

Döbert, H. (2008). Die Bildungsberichterstattung in Deutschland – Oder: Wie können Indikatoren zu Innovationen im Bildungswesen beitragen? In lisum Deutschland, bm:ukk Österreich, & edk Schweiz (Hrsg.), *Bildungsmonitoring, Vergleichsstudien und Innovationen. Von evidenzbasierter Steuerung zur Praxis* (S. 71-91). Berlin: BWV.

Dreßler, J. (2016). Wider eine ökonomische Sicht auf Schule? Die „Neue Steuerung" im Bildungswesen und die „Eigenstruktur des Pädagogischen". In M. Heinrich, & B. Kohlstock (Hrsg.), *Ambivalenzen des Ökonomischen. Analysen zur „Neuen Steuerung" im Bildungssystem* (S. 59-70). Wiesbaden: Springer VS.

Dubs, R. (2010). Methoden und Techniken der Organisationsanalyse. In T. Bohl, W. Helsper, H. G. Holtappels, & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie – Forschungsbefunde – Entwicklungsprozesse – Methodenrepertoire* (S. 481-488). Bad Heilbrunn: Julius Klinkhardt.

Eickelmann, B., Kahnert, J., & Lorenz, R. (2013). Geschlechtsspezifische Fairness im Zentralabitur. Eine Untersuchung im Fach Mathematik. In K. Schwippert, M. Bonsen, & N. Berkemeyer (Hrsg.), *Schul- und Bildungsforschung. Diskussionen, Befunde und Perspektiven. Festschrift für Wilfried Bos* (S. 147-165). Münster et al.: Waxmann.

Emmerich, M., & Maag Merki, K. (2014). Die Entwicklung von Schule. Theorie – Forschung – Praxis. In B. Dippelhofer-Stiem, & S. Dippelhofer (Hrsg.), *Enzyklopädie Erziehungswissenschaft Online*. Weinheim und Basel: Beltz Juventa.

Feldhoff, T., Bischof, L., Emmerich, M., & Radisch, F. (2015). „Was nicht passt, wird passend gemacht!" – Zur Verbindung von Schuleffektivität und Schulentwicklung. In H. J. Abs, T. Brüsemeister, M. Schemmann, & J. Wissinger (Hrsg.), *Governance im Bildungssystem. Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 65-87). Wiesbaden: Springer.

Feldhoff, T., Gromala, L., & Brüsemeister, T. (2014). Organisationales Lernen von Schulen im Kontext datenbasierter Steuerung. In H. G. Holtappels (Hrsg.), *Schulentwicklung und Schulwirksamkeit als Forschungsfeld. Theorieansätze und Forschungserkenntnisse zum schulischen Wandel* (S. 241-257). Münster: Waxmann.

Feldhoff, T., Radisch, F., & Bischof, L. M. (2016). Designs and methods in school improvement research: a systematic review. *Journal of Educational Administration, 54*(2), 209-240. doi:10.1108/JEA-07-2014-0083

Fend, H. (2001). *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung* (2. Auflage). Weinheim & München: Juventa.

Fend, H. (2008a). *Neue Theorie der Schule: Einführung in das Verstehen von Bildungssystemen* (2., durchgesehene Auflage). Wiesbaden: VS.

Fend, H. (2008b). *Schule gestalten: Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS.

Fend, H. (2011). Die Wirksamkeit der Neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. *Zeitschrift für Bildungsforschung, 1*(1), 5-24. doi:10.1007/s35834-011-0003-3

Firestone, W. A., & Schorr, R. Y. (2004). Introduction. In W. A. Firestone, R. Y. Schorr, & L. F. Monfils (Hrsg.), *The Ambiguity of Teaching to the Test. Standards, Assessment, and Educational Reform* (S. 1-17). Mahwah, NJ: Lawrence Erlbaum Associates.

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, & T. Urdan (Hrsg.), *APA Educational Psychology Handbook. Volume 2: Individual Differences and Cultural and Contextual Factors* (S. 471-499). American Psychological Association.

Flaitz, J. (2011). Assessment for Learning: US Perspectives. In R. Berry, & B. Adamson (Hrsg.), *Assessment Reform in Education. Policy and Practice* (S. 33-47). Dordrecht et al.: Springer.

Forte, E. (2010). Examining the Assumptions Underlying the NCLB Federal Accountability Policy on School Improvement. *Educational Psychologist, 45*(2), 76-88.

Frenzel, A. C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R. E. (2009). Emotional Transmission in the Classroom: Exploring the Relationship Between Teacher and Student Enjoyment. *Journal of Educational Psychology, 101*(3), 705-716. doi:10.1037/a0014695

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction, 17*(5), 478-493. doi:10.1016/j.learninstruc.2007.09.001

Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K., & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary educational psychology, 41*, 1-12. doi:10.1016/j.cedpsych.2014.10.006

Fullan, M. (1999). *Die Schule als lernendes Unternehmen. Konzepte für eine neue Kultur der Pädagogik*. Stuttgart: Klett-Cotta.

Fussangel, K., Dizinger, V., Böhm-Kasper, O., & Gräsel, C. (2010). Kooperation, Belastung und Beanspruchung von Lehrkräften an Halb- und Ganztagsschulen. *Unterrichtswissenschaft, 38*(1), 51-67.

Fussangel, K., & Gräsel, C. (2011). Forschung zur Kooperation im Lehrerberuf. In E. Terhart, H. Bennewitz, & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 667-682). Münster et al.: Waxmann.

Gehrmann, A. (2003). *Der Professionelle Lehrer. Muster der Begründung – Empirische Rekonstruktion*. Opladen: Leske + Budrich.

Gehrmann, A. (2007). Zufriedenheit trotz beruflicher Beanspruchungen? Anmerkungen zu den Befunden der Lehrerbelastungsforschung. In M. Rothland (Hrsg.), *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen* (S. 185-203). Wiesbaden: VS.

Gläser-Zikuda, M., & Fuß, S. (2008). Lehrerkompetenzen und Schüleremotionen: Wie nehmen Lernende ihre Lehrkräfte emotional wahr? In M. Gläser-Zikuda, & J. Seifried (Hrsg.), *Lehrerexpertise. Analyse und Bedeutung unterrichtlichen Handelns* (S. 113-142). Münster et al.: Waxmann.

Gläser-Zikuda, M., & Mayring, P. (2003). A qualitative oriented approach to learning emotions at school. In P. Mayring, & C. von Rhöneck (Hrsg.), *Learning Emotions. The Influence of Affective Factors on Classroom Learning* (S. 103-126). Frankfurt am Main et al.: Peter Lang.

Goetz, T., Pekrun, R., Zirngibl, A., Jullien, S., Kleine, M., Vom Hofe, R., & Blum, W. (2004). Leistung und emotionales Erleben im Fach Mathematik – Längsschnittliche Mehrebenenanalysen. *Zeitschrift für pädagogische Psychologie, 18*(3-4), 201-212.

Gogolin, I., Baumert, J., & Scheunpflug, A. (2011). „Transforming education". Large-scale reform projects and their effects–German and International experience. *Zeitschrift für Erziehungswissenschaft, Sonderheft 13*, 1-8. doi:10.1007/s11618-010-0160-5

Goldberg, G. L., & Roswell, B. S. (2010). From Perception to Practice: The Impact of Teachers' Scoring Experience on Performance-Based Instruction and Classroom Assessment. *Educational Assessment, 6*(4), 257-290. doi:10.1207/s15326977ea0604_3

Gomolla, M., & Radtke, F.-O. (2009). *Institutionelle Diskriminierung. Die Herstellung ethnischer Differenz in der Schule* (3. Auflage). Wiesbaden: VS.

Good, T. L., Wiley, C. R. H., & Sabers, D. (2010). Accountability and Educational Reform: A Critical Analysis of Four Perspectives and Considerations for Enhancing Reform Efforts. *Educational Psychologist, 45*(2), 138-148.

Gördel, B.-M. (2015). Der Beitrag der Verwaltungswissenschaft zur Educational Governance-Forschung als interdisziplinäre Wissenschaftsdisziplin. In H. J. Abs, T. Brüsemeister, M. Schemmann, & J. Wissinger (Hrsg.), *Governance im Bildungssystem. Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 143-169). Wiesbaden: Springer VS.

Green, S., Kearbey, J., Wolgemuth, J., Agosto, V., Romano, J., Riley, M., & Frier, A. (2015). Past, Present, and Future of Assessment in Schools: A Thematic Narrative Analysis. *The Qualitative Report, 20*(7), 1111-1124.

Grözinger, G., & Baillet, F. (2015). Gibt es auch beim Abitur eine Noteninflation? Zur Entwicklung der Abiturnoten als Hochschulzugangsberechtigung – Eine Darstellung und Analyse aus soziologischer Perspektive. *Bildung und Erziehung, 68*(4), 473-494.

Haertel, E. (2013). How Is Testing Supposed to Improve Schooling? *Measurement: Interdisciplinary Research & Perspective, 11*(1-2), 1-18. doi:10.1080/15366367.2013.783752

Hahn, J. S. (2014). *Steuerungswirkungen zentraler Vergleichsarbeiten auf den vorgelagerten Unterricht Testcoaching am Beispiel von Lernstand8*. Aachen: Shaker Verlag.

Hallinger, P., & Heck, R. H. (2011). Exploring the journey of school improvement: classifying and analyzing patterns of change in school improvement processes and learning outcomes. *School effectiveness and school improvement, 22*(1), 1-27. doi:10.1080/09243453.2010.536322

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making Sense of Test-Based Accountability in Education*. Santa Monica, CA et al.: RAND.

Hamilton, L. S., Stecher, B. M., Russell, J. L., Marsh, J. A., & Miles, J. (2008). Accountability and teaching practices: school-level actions and teacher responses. In B. Fuller, M. K. Henne, & E. Hannum (Hrsg.), *Strong stakes, weak schools: the benefits and dilemmas of centralized accountability* (S. 31-66). Bingley: Emerald.

Hänlein, A., & Schroeder, W. (2010). Patienteninteressen im deutschen Gesundheitswesen. In U. Clement, J. Nowak, C. Scherrer, & S. Ruß (Hrsg.), *Public Governance und schwache Interessen* (S. 47-61). Wiesbaden: VS.

Haptonstall, K. G. (2010). *An Analysis of the Correlation between Standards-Based, Non-Standards-Based Grading Systems and Achievement as Measured by the Colorado Student Assessment Program (CSAP)*. Ann Arbor, MI: ProQuest LLC.

Hargreaves, A. (1998). The emotional practice of teaching. *Teaching and Teacher Education, 14*(8), 835-854.

Hargreaves, A. (2004). Inclusive and exclusive educational change: emotional responses of teachers and implications for leadership. *School leadership & management, 24*(2), 287-309.

Hargreaves, A. (2005). Educational change takes ages: Life, career and generational factors in teachers' emotional responses to educational change. *Teaching and Teacher Education, 21*(8), 967-983. doi:10.1016/j.tate.2005.06.007

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81-112. doi:10.3102/003465430298487

Heinrich, M., & Kohlstock, B. (2016). *Ambivalenzen des Ökonomischen. Analysen zur „Neuen Steuerung" im Bildungssystem*. Wiesbaden: Springer VS.

Helbig, M., & Nikolai, R. (2015). *Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949.* Bad Heilbrunn: Klinkhardt.

Helmke, A. (2014). Was wissen wir über guten Unterricht? *PADUA, 9*(2), 66-74.

Herman, J. L. (2005). *Making Accountability Work to Improve Student Learning. CSE Report 649*. Retrieved from http://files.eric.ed.gov/fulltext/ED488721.pdf

Herrmann, U. G. (2009). "Alte" und "neue" Steuerung im Bildungssystem. Anmerkungen zu einem bildungshistorisch problematischen Dualismus. In U. Lange, S. Rahn, W. Seitter, & R. Körzel (Hrsg.), *Steuerungsprobleme im Bildungswesen. Festschrift für Klaus Harney* (S. 57-77). Wiesbaden: VS.

Herzog, W. (2016). Durchgriff auf den Lernprozess. Die technologische Reduktion von Schule und Unterricht in der Standardbewegung – am Beispiel der USA. In M. Heinrich, & B. Kohlstock (Hrsg.), *Ambivalenzen des Ökonomischen. Analysen zur „Neuen Steuerung" im Bildungssystem* (S. 109-139). Wiesbaden: Springer VS.

Hochberg, E. D., & Desimone, L. M. (2010). Professional Development in the Accountability Context: Building Capacity to Achieve Standards. *Educational Psychologist, 45*(2), 89-106. doi:10.1080/00461521003703052

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf Schüler- und Schulklassenebene in Primar- und Sekundarstufe*. Münster: Waxmann.

Hofer, S. I. (2015). Studying Gender Bias in Physics Grading: The role of teaching experience and country. *International Journal of Science Education, 37*(17), 2879-2905. doi:10.1080/09500693.2015.1114190

Hofmann, J. (2006). Internet Governance: Eine regulative Idee auf der Suche nach ihrem Gegenstand. In G. Folke Schuppert (Hrsg.), *Governance-Forschung. Vergewisserung über Stand und Entwicklungslinien* (2. Auflage, S. 277-301). Baden-Baden: Nomos.

Holcombe, R., Jennings, J. L., & Koretz, D. (2013). The Roots of Score Inflation: An Examination of Opportunities in Two States' Tests. In G. L. Sunderman (Hrsg.), *Charting reform, achieving equity in a diverse nation* (S. 163-189). Charlotte, NC: Information Age Publishing.

Holmeier, M. (2012a). Bezugsnormorientierung im Unterricht im Kontext zentraler Abiturprüfungen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 237-261). Wiesbaden: Springer VS.

Holmeier, M. (2012b). Vergleichbarkeit der Punktzahlen im schriftlichen Abitur. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 293-324). Wiesbaden: Springer VS.

Holmeier, M. (2013). *Leistungsbeurteilung im Zentralabitur*. Wiesbaden: Springer VS.

Holmeier, M., & Maag Merki, K. (2012). Unterstützung im Unterricht im Kontext der Einführung zentraler Abiturprüfungen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 155-178). Wiesbaden: Springer VS.

Holtappels, H. G. (2010a). Schule als Lernende Organisation. In T. Bohl, W. Helsper, H. G. Holtappels, & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie – Forschungsbefunde – Entwicklungsprozesse – Methodenrepertoire* (S. 99-105). Bad Heilbrunn: Julius Klinkhardt.

Holtappels, H. G. (2010b). Schulentwicklungsforschung. In T. Bohl, W. Helsper, H. G. Holtappels, & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie – Forschungsbefunde – Entwicklungsprozesse – Methodenrepertoire* (S. 26-29). Bad Heilbrunn: Julius Klinkhardt.

Holtappels, H. G. (2014). Schulentwicklung und Schulwirksamkeit. Erkenntnisse aus der Perspektive von Schulentwicklungstheorie und -forschung. In H. G. Holtappels (Hrsg.), *Schulentwicklung und Schulwirksamkeit als Forschungsfeld. Theorieansätze und Forschungserkenntnisse zum schulischen Wandel* (S. 11-47). Münster & New York: Waxmann.

Hoover, R. L. (2014). The Pseudoaccountability of School Reform. Injustice by (False) Proxy. In P. L. Thomas, B. Porfilio, J. Gorlewski, & P. R. Carr (Hrsg.), *Social Context Reform. A Pedagogy of Equity and Opportunity* (S. 49-67). New York & London: Routledge.

Hopkins, D., Ainscow, M., & West, M. (1994). *School Improvement in an Era of Change*. London & New York: Cassell.

Hornberg, S., & Parreira do Amaral, M. (Hrsg.). (2012). *Deregulierung im Bildungswesen*. Münster et al.: Waxmann.

Hospel, V., & Galand, B. (2016). Are both classroom autonomy support and structure equally important for students' engagement? A multilevel analysis. *Learning and Instruction, 41*, 1-10. doi:10.1016/j.learninstruc.2015.09.001

Hugener, I. (2008). *Inszenierungsmuster im Unterricht und Lernqualität. Sichtstrukturen schweizerischen und deutschen Mathematikunterrichts in ihrer Beziehung zu Schülerwahrnehmung und Lernleistung – eine Videoanalyse*. Münster et al.: Waxmann.

Hurrelmann, K. (2006). *Einführung in die Sozialisationstheorie* (9., unveränderte Auflage). Weinheim & Basel: Beltz.

Ingenkamp, K. (2005). *Lehrbuch der pädagogischen Diagnostik* (5., völlig überarbeitete Auflage). Weinheim: Beltz.

Ingenkamp, K. (Hrsg.) (1972). *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte* (3. Auflage). Weinheim: Beltz.

Ivanov, S., Nikolova, R., & Vieluf, U. (2016). G8 vs. G9 im Kohortenvergleich. Lernkontexte und Lernstände zweier Hamburger Abiturjahrgänge. In J. Kramer, M. Neumann, & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte* (S. 81-106). Wiesbaden: Springer VS.

Jäger, D. J. (2012a). Herausforderung Zentralabitur: Unterrichtsinhalte variieren und an Prüfungsthemen anpassen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland*. Wiesbaden: Springer VS.

Jäger, D. J. (2012b). Schulklima, Selbstwirksamkeit und Arbeitszufriedenheit aus Sicht der Lehrpersonen und Schüler/-innen in Hessen und Bremen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 61-89). Wiesbaden: Springer VS.

Jarren, O., & Donges, P. (2004). Staatliche Medienpolitik und die Politik der Massenmedien: Institutionelle und symbolische Steuerung im Mediensystem. In S. Lange, & U. Schimank (Hrsg.), *Governance und gesellschaftliche Integration* (S. 47-63). Wiesbaden: VS.

Jauß, C., & Stark, C. (2004). Kultur und Institution als intervenierende Faktoren in umweltpolitischen Governance-Regimen. In S. Lange, & U. Schimank (Hrsg.), *Governance und gesellschaftliche Integration* (S. 205-225). Wiesbaden: VS.

Jennings, P. A., & Greenberg, M. T. (2009). The Prosocial Classroom: Teacher Social and Emotional Competence in Relation to Student and Classroom Outcomes. *Review of Educational Research, 79*(1), 491-525. doi:10.3102/0034654308325693

Jo, S. H. (2014). Teacher commitment: Exploring associations with relationships and emotions. *Teaching and Teacher Education, 43*, 120-130. doi:10.1016/j.tate.2014.07.004

Jürges, H., & Schneider, K. (2010). Central exit examinations increase performance... but take the fun out of mathematics. *Journal of population economics, 23*(2), 497-517. doi:10.1007/s00148-008-0234-3

Jürges, H., Schneider, K., Senkbeil, M., & Carstensen, C. H. (2009). *Assessment Drives Learning. The Effect of Central Exit Exams on Curricular Knowledge and Mathematical Literacy. CESifo Working Paper No. 2666*. Retrieved from http://www.cesifo-group.de/de/ifoHome/publications/working-papers/CESifoWP/CESifoWPdetails?wp_id=14556274

Jussim, L., Eccles, J., & Madon, S. (1996). Social Perception, Social Stereotypes, and Teacher Expectations: Accuracy and the Quest for the Powerful Self-Fulfilling Prophecy. *Advances in Experimental Social Psychology, 28*, 281-388. doi:10.1016/S0065-2601(08)60240-3

Kahnert, J. (2014). *Das Zentralabitur im Fach Mathematik. Eine empirische Analyse von Abitur- und TIMSS-Daten im Vergleich*. Münster et al.: Waxmann.

Kahnert, J., Eickelmann, B., Lorenz, R., & Bos, W. (2015). Die Steuerungsfunktion von zentralen Abiturprüfungen. Analysen und kontroverse Einschätzungen der Aufgabenschwierigkeit und mögliche Rückkopplungen auf den Unterricht. In H. J. Abs, T. Brüsemeister, M. Schemmann, & J. Wissinger (Hrsg.), *Governance im Bildungssystem. Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (S. 89-115). Wiesbaden: Springer VS.

Kehm, B. M., & Fuchs, M. (2010). Neue Formen der Governance und ihre Folgen für die akademische Kultur und Identität. In U. Clement, J. Nowak, C. Scherrer, & S. Ruß (Hrsg.), *Public Governance und schwache Interessen* (S. 75-94). Wiesbaden: VS.

Kelchtermans, G. (2005). Teachers' emotions in educational reforms: Self-understanding, vulnerable commitment and micropolitical literacy. *Teaching and Teacher Education, 21*(8), 995-1006. doi:10.1016/j.tate.2005.06.009

Klapp Lekholm, A., & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation, 15*(1), 1-23. doi:10.1080/13803610802470425

Klein, E. D. (2016). How do teachers prepare their students for statewide exit exams? A comparison of Finland, Ireland, and the Netherlands. *Journal for educational research online, 8*(2), 31-59.

Klein, E. D., Krüger, M., Kühn, S. M., & van Ackeren, I. (2014). Wirkungen zentraler Abschlussprüfungen im Mehrebenensystem Schule. Eine Zwischenbilanz internationaler und nationaler Befunde und Forschungsdesiderata. *Zeitschrift für Erziehungswissenschaft, 17*(7), 7-33.

Klein, E. D., Kühn, S. M., van Ackeren, I., & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. *Zeitschrift für Pädagogik, 55*(4), 596-621.

Klein, E. D., & van Ackeren, I. (2011). Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels. *Studies in Educational Evaluation, 37*(4), 180-188. doi:10.1016/j.stueduc.2012.01.002

Klemm, K. (2005). Dezentralisierung und Privatisierung im Bildungswesen. In H. G. Holtappels, & K. Höhmann (Hrsg.), *Schulentwicklung und Schulwirksamkeit. Systemsteuerung, Bildungschancen und Entwicklung der Schule* (S. 111-119). Weinheim & München: Juventa.

Klenk, T., & Nullmeier, F. (2004). *Public Governance als Reformstrategie* (2. korrigierte Auflage). Düsseldorf: Hans-Böckler-Stiftung.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos, & R. Lehmann (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 57-128). Opladen: Leske + Budrich.

Klieme, E. (2003). Benotungsmaßstäbe an Schulen: Pädagogische Praxis und institutionelle Bedingungen. Eine empirische Analyse auf der Basis der PISA-Studie. In H. Döbert, B. von Kopp, R. Martini, & M. Weiß (Hrsg.), *Bildung vor neuen Herausforderungen. Historische Bezüge – Rechtliche Aspekte – Steuerungsfragen – Internationale Perspektiven* (S. 195-210). Neuwied: Luchterhand.

Klieme, E., Bürgermeister, A., Harks, B., Blum, W., Leiß, D., & Rakoczy, K. (2010). Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht. Projekt Co2CA. *Zeitschrift für Pädagogik, 56. Beiheft*, 64-74.

Klinger, D. A., & Rogers, W. T. (2011). Teachers' Perceptions of Large-Scale Assessment Programs Within Low-Stakes Accountability Frameworks. *International Journal of Testing, 11*(2), 122-143. doi:10.1080/15305058.2011.552748

Klusmann, U., Kunter, M., & Trautwein, U. (2009). Die Entwicklung des Beanspruchungserlebens bei Lehrerinnen und Lehrern in Abhängigkeit beruflicher Verhaltensstile. *Psychologie in Erziehung und Unterricht, 56*, 200-212.

Koch, S. (2009). Einstellungsmuster von Lehrkräften als Ermöglichung und Begrenzung ‚Neuer Steuerung' – Eine empirische Rekonstruktion. In U. Lange, S. Rahn, & R. Körzel (Hrsg.), *Steuerungsprobleme im Bildungswesen. Festschrift für Klaus Harney* (S. 117-135). Wiesbaden: VS.

Köck, W. (2006). Governance in der Umweltpolitik. In G. Folke Schuppert (Hrsg.), *Governance-Forschung. Vergewisserung über Stand und Entwicklungslinien* (2. Auflage, S. 322-345). Baden-Baden: Nomos.

Kohler, B., & Wacker, A. (2013). Das Angebots-Nutzungs-Modell. Überlegungen zu Chancen und Grenzen des derzeit prominentesten Wirkmodells der Schul- und Unterrichtsforschung. *Die deutsche Schule, 105*(3), 241-257.

Koretz, D. (2008a). *Measuring up. What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Koretz, D. (2008b). Test-based Educational Accountability. Research Evidence and Implications. *Zeitschrift für Pädagogik, 54*(6), 777-790.

Koretz, D. (2011). Lessons from test-based education reform in the U.S. *Zeitschrift für Erziehungswissenschaft, Sonderheft 13*, 9-23.

Krause, A., Dorsenmagen, C., & Alexander, T. (2011). Belastung und Beanspruchung im Lehrerberuf – Arbeitsplatz- und bedingungsbezogene Forschung. In E. Terhart, H. Bennewitz, & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 788-813). Münster et al.: Waxmann.

Krüger, M. (2015). *Aufgabenkultur in zentralen Abschlussprüfungen. Exploration und Deskription naturwissenschaftlicher Aufgabenstellungen im internationalen Vergleich*. Münster & New York: Waxmann.

Kühn, S. M. (2010). *Steuerung und Innovation durch Abschlussprüfungen?* Wiesbaden: VS.

Kühn, S. M. (2011). Exploring the use of statewide exit exams to spread innovation—The example of Context in science tasks from an international comparative perspective. *Studies in Educational Evaluation, 37*(4), 189-195. doi:10.1016/j.stueduc.2012.01.003

Kühn, S. M. (2012). Zentrale Abiturprüfungen im nationalen und internationalen Vergleich mit besonderer Perspektive auf Bremen und Hessen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 27-44). Wiesbaden: Soringer VS.

Kühn, S. M. (2016). Öffnung des Gymnasiums durch die Wiedereinführung von G9? Herausforderungen und Befunde im Kontext der aktuellen Heterogenitätsdebatte. In J. Kramer, M. Neumann, & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte* (S. 107-128). Wiesbaden: Springer VS.

Kühn, S. M., & Racherbäumer, K. (2013). Standardisierung und/oder Individualisierung? Empirische Befunde zur Umsetzung von Maßnahmen zur individuellen Förderung im Kontext zentraler Abschlussprüfungen. *Unterrichtswissenschaft, 41*(2), 172-189.

Kussau, J., & Brüsemeister, T. (2007). Educational Governance: Zur Analyse der Handlungskoordination im Mehrebenensystem der Schule. In H. Altrichter, T. Brüsemeister, & J. Wissinger (Hrsg.), *Educational Governance. Handlungskoordination und Steuerung im Bildungssystem* (S. 15-54). Wiesbaden: VS.

Kyriakides, L. (2005). Extending the Comprehensive Model of Educational Effectiveness by an Empirical Investigation. *School effectiveness and school improvement, 16*(2), 103-152. doi:10.1080/09243450500113936

Kyriakides, L. (2008). Testing the validity of the comprehensive model of educational effectiveness: a step towards the development of a dynamic model of effectiveness. *School effectiveness and school improvement, 19*(4), 429-446. doi:10.1080/09243450802535208

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*, 143-152. doi:10.1016/j.tate.2013.07.010

Kyriakides, L., Creemers, B. P. M., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: implications for theory and research. *British Educational Research Journal, 36*(5), 807-830. doi:10.1080/01411920903165603

Lange, S., & Schimank, U. (2004). *Governance und gesellschaftliche Integration*. Wiesbaden: VS.

Lange, U., Rahn, S., Seitter, W., & Körzel, R. (2009). Zur Einführung: Steuerungsprobleme im Bildungswesen. In U. Lange, S. Rahn, W. Seitter, & R. Körzel (Hrsg.), *Steuerungsprobleme im Bildungswesen. Festschrift für Klaus Harney* (S. 9-15). Wiesbaden: VS.

Lazarides, R., & Watt, H. M. G. (2015). Girls' and boys' perceived mathematics teacher beliefs, classroom learning environments and mathematical career intentions. *Contemporary educational psychology, 41*, 51-61. doi:10.1016/j.cedpsych.2014.11.005

Ledergerber, C. (2015). *Unterrichtskommunikation und motivational-emotionale Aspekte des Lernens. Eine videobasierte Analyse im Mathematikunterricht*. Münster & New York: Waxmann.

Lee, J. (2008). Is Test-Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies. *Review of Educational Research, 78*(3), 608-644. doi:10.3102/0034654308324427

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Wiesbaden: VS.

LISUM Deutschland, bm:ukk Österreich, & EDK Schweiz (Hrsg.). (2008). *Bildungsmonitoring, Vergleichsstudien und Innovationen. Von evidenzbasierter Steuerung zur Praxis*. Berlin: BWV.

Lorenz, R. (2016). Does gender make a difference? Gender-related fairness of high-stakes testing in A-level examinations in English as foreign language in the German state of North Rhine-Westphalia in the context of Educational Governance. *Journal for educational research online, 8*(2), 10-30.

Lüsebrink, I. (2002). Unsicherheit als Herausforderung. Ein Beitrag zur Professionalisierung des LehrerInnenberufs. *Die deutsche Schule, 94*(1), 39-49.

Maag Merki, K. (2008). Die Einführung des Zentralabiturs in Bremen. Eine Fallanalyse. *Die deutsche Schule, 100*(3), 357-366.

Maag Merki, K. (2012a). Die Leistungen der Gymnasiastinnen und Gymnasiasten in Mathematik und Englisch. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 263-292). Wiesbaden: Springer VS.

Maag Merki, K. (2012b). Forschungsfragen und theoretisches Rahmenmodell. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 11-25). Wiesbaden: Springer VS.

Maag Merki, K. (2012c). *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland*. Wiesbaden: Springer VS.

Maag Merki, K. (2012d). Zentrale Prüfungen – empirische Evidenzen der Effekte der Einführung zentraler Abiturprüfungen auf Motivation und Emotion der Schüler/innen. In A. Wacker, U. Maier, & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (S. 247-275). Wiesbaden: Springer VS.

Maag Merki, K. (2014). Das quasi-experimentelle Design in der Educational Governance-Forschung? Herausforderungen, Möglichkeiten und Grenzen am Beispiel der Analyse der Wirksamkeit der Einführung zentraler Abiturprüfungen. In K. Maag Merki, R. Langer, & H. Altrichter (Hrsg.), *Educational Governance als Forschungsperspektive. Strategien. Methoden. Ansätze* (2., erweiterte Auflage, S. 51-83). Wiesbaden: Springer VS.

Maag Merki, K. (2016). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (2., überarbeitete und aktualisierte Auflage, S. 151-181). Wiesbaden: Springer VS.

Maag Merki, K., & Emmerich, M. (2011). Schulexterne Steuerungsinstrumente der Schulentwicklung. In H. Altrichter, & C. Helm (Hrsg.), *Akteure & Instrumente der Schulentwicklung* (S. 151-168). Baltmannsweiler: Schneider.

Maag Merki, K., & Holmeier, M. (2008). Die Implementation zentraler Abiturprüfungen. Erste Ergebnisse zu den Effekten der Einführung auf das schulische Handeln der Lehrpersonen. In E.-M. Lankes (Hrsg.), *Pädagogische Professionalität als Gegenstand empirischer Forschung* (S. 233-243). Münster et al.: Waxmann.

Maag Merki, K., & Holmeier, M. (2012). Selbstreguliertes Lernen der Schülerinnen und Schüler in der Vorbereitung auf das Abitur. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 325-352). Wiesbaden: Springer VS.

Maag Merki, K., & Holmeier, M. (2015). Comparability of semester and exit exam grades: long-term effect of the implementation of state-wide exit exams. *School effectiveness and school improvement, 26*(1), 57-74. doi:10.1080/09243453.2013.861353

Maag Merki, K., Klieme, E., & Holmeier, M. (2008). Unterrichtsgestaltung unter den Bedingungen zentraler Abiturprüfungen. Differenzielle Analysen auf Schulebene mittels Latent Class Analysen. *Zeitschrift für Pädagogik, 54*(6), 791-808.

Maag Merki, K., & Oerke, B. (2012). Methodische Grundlagen der Studie. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 45-61). Wiesbaden: Springer VS.

Maag Merki, K., & Oerke, B. (2016). Long-term effects of the implementation of state-wide exit exams: a multilevel regression analysis of mediation effects of teaching practices on students' motivational orientations. *Educational Assessment, Evaluation and Accountability*, 1-32. doi:10.1007/s11092-016-9244-y

Maag Merki, K., & Werner, S. (2013). Schulentwicklungsforschung. Aktuelle Schwerpunkte und zukünftige Forschungsperspektiven. *Die deutsche Schule, 105*(3), 295-304.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology, 1*, 86-92. doi:10.1027/1614-1881.1.3.86

Maaz, K., Baeriswyl, F., & Trautwein, U. (2011). *Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Eine Studie im Auftrag der Vodafone Stiftung Deutschland.* Retrieved from https://www.vodafone-stiftung.de/alle_publikationen.html?&tx_newsjson_pi1%5BshowUid%5D=44&cHash=77c9bde76786dbdff61bdb720d3dff8c

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high stakes testing on minority students: evidence from 100 yeares of test data. In G. Orfield, & M. Kornhaber (Hrsg.), *Raising standards or raising barriers? Inequality and high stakes testing in public education* (S. 1-49). New York: The Century Foundation.

Madaus, G. F., Russell, M. K., & Higgins, J. (2009). *The paradoxes of high stakes testing. How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing.

Mansell, W. (2011). Improving exam results, but to what end? The limitations of New Labour's control mechanism for schools: assessment-based accountability. *Journal of Educational Administration and History, 43*(4), 291-308. doi:10.1080/00220620.2011.606896

Maritzen, N. (2008). Bildungsmonitoring – Systeminnovation zur Sicherung von Qualitätsstandards. In LISUM Deutschland, bm:ukk Österreich, & EDK Schweiz (Hrsg.), *Bildungsmonitoring, Vergleichsstudien und Innovationen. Von evidenzbasierter Steuerung zur Praxis* (S. 109-124). Berlin: BWV.

Maritzen, N. (2011). On the advantage and disadvantage of educational monitoring in a federal system. *Zeitschrift für Erziehungswissenschaft, Sonderheft 13*, 117-135.

Maué, E. (2013). Vergleichbarkeit von Abiturnoten – eine Fiktion? Längerfristige Effekte der Implementation zentraler Abiturprüfungen in Bremen. In J. Asdonk, S. U. Kuhnen & P. Bornkessel (Hrsg.). *Von der Schule zur Hochschule. Analysen, Konzeptionen und Gestaltungsperspektiven des Übergangs* (S. 114-128). Münster: Waxmann.

Maué, E. (2016). Achievement—and what else? The standardisation of semester grades due to the implementation of state-wide exit examinations. *Studies in Educational Evaluation, 51*, 42-54. doi:10.1016/j.stueduc.2016.09.003

Maué, E. (2017). Die Implementation zentraler Abiturprüfungen und deren potentielle Auswirkungen auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg von Schülerinnen und Schülern. *Zeitschrift für Pädagogik, 63*(6), 803-826.

Maué, E., Maag Merki, K., & Oerke, B. (2012). Emotionales Erleben des Zentralabiturs von Lehrpersonen in Bremen. Längerfristige Effekte der Implementation zentraler Abiturprüfungen. In S. Hornberg, & M. Parreira do Amaral (Hrsg.), *Deregulierung im Bildungswesen* (S. 109-130). Münster et al.: Waxmann.

Maynitz, R. (2009). *Über Governance. Institutionen und Prozesse politischer Regelung.* Frankfurt am Main & New York: Campus.

Meijer, J. (2007). Correlates of student stress in secondary education. *Educational research, 49*(1), 21-35. doi:10.1080/00131880701200708

Meredith, C., Moolenaar, N. M., Struyve, C., Vandecandelaere, M., Gielen, S., & Kyndt, E. (2017). The measurement of collaborative culture in secondary schools: An informal subgroup approach. *Frontline Learning Research, 5*(2), 34-45. doi:10.14786/flr.v5i2.283

Meyer-Hesemann, W. (2010). Bildungsreform im Bildungsföderalismus. Ein zweigliedriges Schulsystem für Deutschland ist möglich. Anmerkungen aus gegebenem Anlass. *Die deutsche Schule, 102*(1), 86-90.

Monfils, L. F., Firestone, W. A., Hicks, J. E., Martinez, M. C., Schorr, R. Y., & Camilli, G. (2004). Teachig to the Test. In W. A. Firestone, R. Y. Schorr, & L. F. Monfils (Hrsg.), *The Ambiguity of Teaching to the Test. Standards, Assessment, and Educational Reform* (S. 37-61). Mahwah, NJ: Lawrence Erlbaum Associates.

Muders, S. (2016). Pfadabhängigkeit von Schulen aus organisationstheoretischer Perspektive. In M. Heinrich, & B. Kohlstock (Hrsg.), *Ambivalenzen des Ökonomischen. Analysen zur „Neuen Steuerung" im Bildungssystem* (S. 245-260). Wiesbaden: Springer VS.

Muijs, D., Campbell, J., & Kyriakides, L. (2005). Making the case for differentiated teacher effectiveness. An overview of research in four key areas. *School effectiveness and school improvement, 16*(1), 51-70.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School effectiveness and school improvement, 25*(2), 231-256.

Müller-Benedict, V., & Grözinger, G. (2017). *Noten an Deutschlands Hochschulen. Analysen zur Vergleichbarkeit von Examensnoten 1960 bis 2013*. Wiesbaden: Springer VS. doi:10.1007/978-3-658-15801-9

Muller, C. L. (2015). Measuring School Contexts. *AERA Open, 1*(4). doi:10.1177/2332858415613055

Munthe, E. (2001). Measuring Teacher Certainty. *Scandinavian Journal of Educational Research, 45*(2), 167-181.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (7th Edition). Los Angeles, CA: Muthén & Muthén.

Natriello, G. (2009). High Stakes Testing and Teaching to the Test. In L. J. Saha, & A. G. Dworkin (Hrsg.), *International Handbook of Research on Teachers and Teaching* (S. 1101-1111). New York: Springer.

Neumann, M. (2014). Das Abitur in Deutschland – Aktuelle Entwicklungen und Herausforderungen im Überblick. In F. Eberle, B. Schneider-Taylor, & D. Bosse (Hrsg.), *Abitur und Matura zwischen Hochschulvorbereitung und Berufsorientierung* (S. 245-259). Wiesbaden: Springer VS.

Neumann, M., Nagy, G., Trautwein, U., & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen. Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. *Zeitschrift für Erziehungswissenschaft, 12*(4), 691-714. doi:10.1007/s11618-009-0099-6

Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. *Studies in Educational Evaluation, 37*(4), 206-217. doi:10.1016/j.stueduc.2012.02.002

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1), 1-175.

OECD. (2012). *PISA. Grade Expectations: How Marks and Education Policies Shape Students' Ambitions*. OECD Publishing. doi:10.1787/9789264187528-en

Oelkers, J., & Reusser, K. (2008). *Qualität entwickeln – Standards sichern – mit Differenz umgehen. Bildungsforschung Band 27.* Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Oerke, B. (2012a). Auseinandersetzung der Lehrpersonen mit der Einführung des Zentralabiturs: Stages of Concern. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 203-232). Wiesbaden: Springer VS.

Oerke, B. (2012b). Emotionaler Umgang von Lehrkräften und Schüler/-innen mit dem Zentralabitur: Unsicherheit, Leistungsdruck und Leistungsattributionen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 115-149). Wiesbaden: Springer VS.

Oerke, B., Maag Merki, K., Holmeier, M., & Jäger, D. J. (2011). Changes in student attributions due to the implementation of central exit exams. *Educational Assessment, Evaluation and Accountability, 23*(3), 223-241. doi:10.1007/s11092-011-9121-7

Oerke, B., Maag Merki, K., Maué, E., & Jäger, D. J. (2013). Zentralabitur und Themenvarianz im Unterricht. Lohnt sich Teaching-to-the-Test? In D. Bosse, F. Eberle, & B. Schneider-Taylor (Hrsg.), *Standardisierung in der gymnasialen Oberstufe* (S. 27-49). Wiesbaden: Springer VS.

Oevermann, U. (2002). Professionalisierungsbedürftigkeit und Preofessionalisiertheit pädagogischen Handelns. In M. Kraul, W. Marotzki, & C. Schweppe (Hrsg.), *Biographie und Profession* (S. 19-63). Bad Heilbrunn: Julius Klinkhardt.

Oevermann, U. (2008). Profession contra Organisation? Strukturtheoretische Perspektiven zum Verhältnis von Organisation und Profession in der Schule. In W. Helsper, S. Busse, M. Hummrich, & R.-T. Kramer (Hrsg.), *Pädagogische Professionalität in Organisationen. Neue Verhältnisbestimmungen am Beispiel der Schule* (S. 55-77). Wiesbaden: VS.

Paeplow, C. G. (2008). *Middle School Grading: Wake County Public School System (WCPSS) 2006-07 and 2007-08. E&R Report No. 08.16*. Retrieved from Raleigh, NC: http://www.wcpss.net/results/reports/2008/0816ms_grading2008.pdf

Paeplow, C. G. (2011). *Easy as 1, 2, 3: Exploring the Implementation of Standards-Based Grading in Wake County Elementary Schools.* North Carolina State University. Retrieved from http://repository.lib.ncsu.edu/ir/bitstream/1840.16/7242/1/etd.pdf

Parreira do Amaral, M. (2012). Governance und Deregulierung von Bildung. Regimetheoretische Überlegungen zu einem internationalen Trend. In S. Hornberg, & M. Parreira do Amaral (Hrsg.), *Deregulierung im Bildungswesen* (S. 71-92). Münster et al.: Waxmann.

Pedulla, J., Abrams, L. M., Madaus, G., Russel, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy, Lynch School of Education, Boston College.

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review, 18*(4), 315-341.

Penninckx, M., Quintelier, A., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2017). Delphi study on standardized systems to monitor student learning outcomes in Flanders: mechanisms for building trust and/or control? *Studia paedagogica, 22*(2), 9-31. doi:10.5817/SP2017-2-2

Penuel, W. R., Frank, K. A., Sun, M., & Kim, C. M. (2012). Teachers' Social Capital and the Implementation of Schoolwide Reforms. In S. Kelly (Hrsg.), *Assessing Teacher Quality. Understanding Teacher Effects on Instruction and Achievement* (S. 183-200). New York & London: Teachers College, Columbia University.

Piopiunik, M., Schwerdt, G., & Wößmann, L. (2014). Zentrale Abschlussprüfungen, Signalwirkung von Abiturnoten und Arbeitsmarkterfolg in Deutschland. *Zeitschrift für Erziehungswissenschaft, 17*(1), 35-60.

Prodromou, L. (1995). The backwash effect: from testing to teaching. *ELT journal, 49*, 13-25. doi:10.1093/elt/49.1.13

Putwain, D. (2008). Do examinations stakes moderate the test anxiety–examination performance relationship? *Educational Psychology, 28*(2), 109-118.

Rakoczy, K. (2006). Motivationsunterstützung im Mathematikunterricht. Zur Bedeutung von Unterrichtsmerkmalen für die Wahrnehmung von Schülerinnen und Schülern. *Zeitschrift für Pädagogik, 52*(6), 822-843.

Rakoczy, K., Klieme, E., Bürgermeister, A., & Harks, B. (2008). The Interplay Between Student Evaluation and Instruction Grading and Feedback in Mathematics Classrooms. *Zeitschrift für Psychologie, 216*(2), 111-124. doi:10.1027/0044-3409.216.2.111

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International, Inc.

Ravitch, D. (2010). *The death and life of the great American school system. How testing and choice are undermining education*. New York: Basic Books.

Reardon, S. F., Atteberry, A., Arshan, N., & Kurlaender, M. (2009). *Effects of the California High School Exit Exam on Student Persistence, Achievement, and Graduation. Working Paper 2009-12*. Retrieved from http://web.stanford.edu/group/cepa/workingpapers/WORKING_PAPER_2009_12.pdf

Resh, N. (2009). Justice in grades allocation: teachers' perspective. *Social Psychology of Education, 12*(3), 315-325. doi:10.1007/s11218-008-9073-z

Resh, N. (2010). Sense of justice about grades in school: is it stratified like academic achievement? *Social Psychology of Education, 13*(3), 313-329. doi:10.1007/s11218-010-9117-z

Reusser, K., & Halbheer, U. (2008). Die Implementation von Bildungsstandards als Anstoß zur Qualitätsentwicklung in Schule und Unterricht. In LISUM Deutschland, bm:ukk Österreich, & EDK Schweiz (Hrsg.), *Bildungsmonitoring, Vergleichsstudien und Innovationen. Von evidenzbasierter Steuerung zur Praxis* (S. 125-138). Berlin: BWV.

Reusser, K., & Pauli, C. (2010). Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht: Einleitung und Überblick. In K. Reusser, C. Pauli, & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 9-32). Münster et al.: Waxmann.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom Emotional Climate, Student Engagement, and Academic Achievement. *Journal of Educational Psychology, 104*(3), 700-712.

Reynolds, D. (2005). School Effectiveness: Past, Present and Future Directions. In H. G. Holtappels, & K. Höhmann (Hrsg.), *Schulentwicklung und Schulwirksamkeit. Systemsteuerung, Bildungschancen und Entwicklung der Schule* (S. 11-25). Weinheim & München: Juventa.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): a state-of-the-art review. *School effectiveness and school improvement, 25*(2), 197-230. doi:10.1080/09243453.2014.885450

Reynolds, D., Teddlie, C., Chapman, C., & Stringfield, S. (2015). Effective school processes. In C. Chapman, D. Muijs, & D. Reynolds (Hrsg.), *The Routledge International Handbook of Educational Effectiveness and Improvement: Research, policy, and practice* (S. 77-99). Florence: Routledge.

Reynolds, W. M. (2014). Reforming the Schooling of Neoliberal Perpetual Zombie Desire. In P. L. Thomas, B. Porfilio, J. Gorlewski, & P. R. Carr (Hrsg.), *Social Context Reform. A Pedagogy of Equity and Opportunity* (S. 33-48). New York, NY: Taylor & Francis.

Roderick, M., Jacob, B. A., & Bryk, A. S. (2002). The Impact of High-Stakes Testing in Chicago on Student Achievement in Promotional Gate Grades. *Educational evaluation and policy analysis, 24*(4), 333-357.

Rogers, E. M. (2003). *Diffusion of innovations* (5th). New York: Free Press.

Rolff, H.-G. (1993). *Wandel durch Selbstorganisation. Theoretische Grundlagen und praktische Hinweise für eine bessere Schule.* Weinheim & München: Juventa.

Rolff, H.-G. (2010). Schulentwicklung als Trias von Organisations-, Unterrichts- und Personalentwicklung. In T. Bohl, W. Helsper, H. G. Holtappels, & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie – Forschungsbefunde – Entwicklungsprozesse – Methodenrepertoire* (S. 29-36). Bad Heilbrunn: Julius Klinkhardt.

Rolff, H.-G. (2013). *Schulentwicklung kompakt. Modelle, Instrumente, Perspektiven*. Weinheim: Beltz.

Roos, A.-L., Bieg, M., Goetz, T., Frenzel, A. C., Taxer, J., & Zeidner, M. (2015). Experiencing more mathematics anxiety than expected? Contrasting trait and state anxiety in high achieving students. *High Ability Studies, 26*(2), 245-258. doi:10.1080/13598139.2015.1095078

Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the Classroom. Teacher Expectation and Pupils' Intellectual Development* (erweiterte Auflage). New York: Irvington.

Roßnagel, A., & von Wangenheim, G. (2010). Schwache Interessen in der Selbstregulierung im Umweltrecht. In U. Clement, J. Nowak, C. Scherrer, & S. Ruß (Hrsg.), *Public Governance und schwache Interessen* (S. 127-139). Wiesbaden: VS.

Rost, D. H., & Schermer, F. J. (1987). Emotion and Cognition in Coping with Test Anxiety. *Communication & Cognition, 20*(2/3), 225-244.

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: teacher expectations and academic outcomes. *British Journal of Educational Psychology, 76*(3), 429-444. doi:10.1348/000709905X53589

Rubin, D. B. (1987). *Mutliple Imputation for nonresponse in surveys*. New York: Wiley.

Rudow, B. (1994). *Die Arbeit des Lehrers. Zur Psychologie der Lehrertätigkeit, Lehrerbelastung und Lehrergesundheit*. Bern et al.: Hans Huber.

Rürup, M. (2011). Innovationen im Bildungswesen: Begriffliche Annäherungen an das Neue. *Die deutsche Schule, 103*(1), 9-23.

Rürup, M., Fuchs, H.-W., & Weishaupt, H. (2016). Bildungsberichterstattung – Bildungsmonitoring. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (2., überarbeitete und aktualisierte Auflage, S. 411-437). Wiesbaden: Springer VS.

Rürup, M., & Heinrich, M. (2007). Schulen unter Zugzwang – Die Schulautonomiegesetzgebung der deutschen Länder als Rahmen der Schulentwicklung. In H. Altrichter, T. Brüsemeister, & J. Wissinger (Hrsg.), *Educational Governance. Handlungskoordination und Steuerung im Bildungssystem* (S. 157-183). Wiesbaden: VS.

Rustemeyer, D. (2009). Anarchie im Büro? Organisation als Formen multipler Rationalität. In U. Lange, S. Rahn, W. Seitter, & R. Körzel (Hrsg.), *Steuerungsprobleme im Bildungswesen. Festschrift für Klaus Harney* (S. 35-56). Wiesbaden: VS.

Ryan, K. E., Ryan, A. M., Arbuthnot, K., & Samuels, M. (2007). Students' Motivation for Standardized Math Exams. *Educational Researcher, 36*(1), 5-13.

Sahner, B. (2008). Professionalitätsgestaltungen von Lehrerinnen und Lehrern und ihre Schulentwicklungskompetenz. In G. Breidenstein, & F. Schütze (Hrsg.), *Paradoxien in der Reform der Schule. Ergebnisse qualitativer Sozialforschung* (S. 261-272). Wiesbaden: VS.

Schaarschmidt, U., & Kieschke, U. (2007). Beanspruchungsmuster im Lehrerberuf. Ergebnisse und Schlussfolgerungen aus der Potsdamer Lehrerstudie. In M. Rothland (Hrsg.), *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen* (S. 81-98). Wiesbaden: VS.

Scheerens, J. (1990). School Effectiveness Research and the Development of Process Indicators of School Functioning. *School Effectiveness and School Improvement, 1*(1), 61-80. doi: 10.1080/0924345900010106

Scheerens, J. (1992). *Effective Schooling. Research, Theory and Practice*. London & New York: Cassell.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating Statistical Power and Required Sample Sizes for Organizational Research Using Multilevel Modeling. *Organizational Research Methods, 12*(2), 347-367. doi:10.1177/1094428107308906

Schildkamp, K., Rekers-Mombarg, L. T. M., & Harms, T. J. (2012). Student group differences in examination results and utilization for policy and school development. *School effectiveness and school improvement, 23*(2), 229-255.

Schmidt, M., & Datnow, A. (2005). Teachers' sense-making about comprehensive school reform: The influence of emotions. *Teaching and Teacher Education, 21*, 949-965. doi:10.1016/j.tate.2005.06.006

Schräpler, J.-P., & Weishaupt, H. (2013). Auswirkungen des Zentralabiturs auf den Abiturerfolg an Gymnasien und Gesamtschulen in Nordrhein-Westfalen. In N. McElvany, & H. G. Holtappels (Hrsg.), *Empirische Bildungsforschung. Theorien, Methoden, Befunde und Perspektiven* (S. 249-266). Münster et al.: Waxmann.

Schratz, M., & Steiner-Löffler, U. (1999). *Die Lernende Schule. Arbeitsbuch pädagogische Schulentwicklung* (2., korrigierte Auflage). Weinheim & Basel: Beltz.

Schraw, G. (2010). No School Left Behind. *Educational Psychologist, 45*(2), 71-75. doi:10.1080/00461521003720189

Schumacher, C. (2016). *Prüfungsangst in der Schule. Ursachen, Bewältigung und Folgen am Beispiel zentraler Abschlussprüfung*. Münster & New York: Waxmann.

Schütze, F., & Breidenstein, G. (2008). Überlegungen zum paradoxen Charakter von Schulreformprozessen – eine Einleitung. In G. Breidenstein, & F. Schütze (Hrsg.), *Paradoxien in der Reform der Schule. Ergebnisse qualitativer Sozialforschung* (S. 9-23). Wiesbaden: VS.

Schwarzer, R. (2000). *Streß, Angst und Handlungsregulation* (4., überarbeitete Auflage). Stuttgart: Kohlhammer.

Scott, D. (2011). Assessment Reform: High-Stakes Testing and Knowing the Contents of Other Minds. In R. Berry, & B. Adamson (Hrsg.), *Assessment Reform in Education. Policy and Practice* (S. 155-163). Dordrecht et al.: Springer.

Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie. Integration von Struktur- und Prozessparadigma. *Zeitschrift für Pädagogik, 60*(6), 850-866.

Seifried, J. (2009). *Unterricht aus der Sicht von Handelslehrern*. Frankfurt am Main: Peter Lang.

Seipp, B. (1990). *Angst und Leistung in Schule und Hochschule. Eine Meta-Analyse*. Frankfurt am Main et al.: Lang.

Smyth, J. (2011). The disaster of the ´self-managing school´ – genesis, trajectory, undisclosed agenda, and effects. *Journal of Educational Administration and History, 43*(2), 95-117. doi:10.1080/00220620.2011.560253

Solórzano, R. W. (2008). High Stakes Testing: Issues, Implications, and Remedies for English Language Learners. *Review of Educational Research, 78*(2), 260-329. doi:10.3102/003465430831784

Soltau, A., & Mienert, M. (2010). Unsicherheit im Lehrerberuf als Ursache mangelnder Lehrerkooperation? Eine Systematisierung des aktuellen Forschungsstandes auf Basis des transaktionalen Stressmodells. *Zeitschrift für Pädagogik, 56*(5), 761-778.

Specht, W. (2008). Innovation durch Evaluation? Entstehung und Umsetzung von Innovationen im Bildungssystem als Konsequenz aus Bildungsmonitoring, Bildungsberichterstattung und vergleichenden Schulleistungsstudien – Möglichkeiten und Grenzen aus österreichischer Sicht. In LISUM Deutschland, bm:ukk Österreich, & EDK Schweiz (Hrsg.), *Bildungsmonitoring, Vergleichsstudien und Innovationen. Von evidenzbasierter Steuerung zur Praxis* (S. 41-52). Berlin: BWV.

Speyer, B. (2006). Governance internationaler Finanzmärkte – zur Erklärung der Polymorphie. In G. Folke Schuppert (Hrsg.), *Governance-Forschung. Vergewisserung über Stand und Entwicklungslinien* (2. Auflage, S. 302-321). Baden-Baden: Nomos.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy Implementation and Cognition: Reframing and Refocusing Implementation Research. *Review of Educational Research, 72*(3), 387-431.

Stanat, P., Becker-Mrotzek, M., Blum, W., & Tesch, B. (2016). Vergleichbarkeit in der Vielfalt. Bildungsstandards der Kultusministerkonferenz für die Allgemeine Hochschulreife. In J. Kramer, M. Neumann, & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte* (S. 29-58). Wiesbaden: Springer VS.

Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2008). *Vereinbarung über Einheitliche Prüfungsanforderungen in der Abiturprüfung*. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2008/2008_10_24-VB-EPA.pdf

Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2013). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II*. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1972/1972_07_07-Vereinbarung-Gestaltung-Sek2.pdf

Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2016). *Die Abiturprüfung in der gymnasialen Oberstufe*. Retrieved from http://www.kmk.org/bildung-schule/allgemeine-bildung/sekundarstufe-ii-gymnasiale-oberstufe/abitur/abiturpruefung-in-der-gymnasialen-oberstufe.html

Stevens, P. A. J., & Görgöz, R. (2010). Exploring the importance of institutional contexts for the development of ethnic stereotypes: a comparison of schools in Belgium and England. *Ethnic and Racial Studies, 33*(8), 1350-1371. doi:10.1080/01419870903219243

Sunderman, G. L. (2013). *Charting reform, achieving equity in a diverse nation*. Charlotte, NC: Information Age Publishing.

Teddlie, C., & Reynolds, D. (2000). *The International Handbook of School Effectiveness Research*: Falmer Press.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99*(2), 253-273. doi:10.1037/0022-0663.99.2.253

Terhart, E. (2008). Giving marks – constructing differences. Explorations in the micro-politics of selection in schools. In H.-H. Krüger, W. Helsper, G. Foljanty-Jost, R.-T. Kramer, & M. Hummrich (Hrsg.), *Family, School, Youth Culture. International Perspectives of Pupil Research* (S. 151-161). Frankfurt am Main et al.: Peter Lang.

Terhart, E. (2010). Schulentwicklung und Lehrerkompetenzen. In T. Bohl, W. Helsper, H. G. Holtappels, & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie – Forschungsbefunde – Entwicklungsprozesse – Methodenrepertoire* (S. 237-241). Bad Heilbrunn: Julius Klinkhardt.

Thiel, F. (2016). *Interaktion im Unterricht. Ordnungsmechanismen und Stördynamiken*. Opladen & Toronto: Barbara Budrich.

Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., & Peetsma, T. T. D. (2012). Building school-wide capacity for improvement: the role of leadership, school organizational conditions, and teacher factors. *School effectiveness and school improvement, 23*(4), 441-460. doi:10.1080/09243453.2012.678867

Thorsen, C. (2012). Dimensions of Norm-Referenced Compulsory School Grades and their Relative Importance for the Prediction of Upper Secondary School Grades. *Scandinavian Journal of Educational Research*(iFirst Article), 1-20. doi:10.1080/00313831.2012.705322

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation, 18*(2), 153-172. doi:10.1080/13803611.2012.659929

Tierney, R. D., Simon, M., & Charland, J. (2011). Being Fair: Teachers' Interpretations of Principles for Standards-Based Grading. *The Educational forum, 75*(3), 210-227. doi:10.1080/00131725.2011.577669

Torrance, H., & Pryor, J. (2008). The social construction of success and failure in classroom assessment in England. In H.-H. Krüger, W. Helsper, G. Foljanty-Jost, R.-T. Kramer, & M. Hummrich (Hrsg.), *Family, School, Youth Culture. Internationale Perspectives of Pupil Research* (S. 219-237). Frankfurt am Main et al.: Peter Lang.

Trautwein, U., Köller, O., Lehmann, R., & Lüdtke, O. (2007). *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten*. Münster: Waxmann.

Trouilloud, D. O., Sarrazin, P. G., Martinek, T. J., & Guillet, E. (2002). The influence of teacher expectations on student achievement in physical education classes: Pygmalion revisited. *European journal of social psychology, 32*(5), 591-607. doi:10.1002/ejsp.109

Tyack, D., & Tobin, W. (1994). The "Grammar" of Schooling: Why Has It Been So Hard to Change? *American educational research journal, 31*(3), 453-479.

Urhahne, D. (2015). Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion. *Teaching and Teacher Education, 45*, 73-82. doi:10.1016/j.tate.2014.09.006

Vähäsantanen, K. (2015). Professional agency in the stream of change: Understanding educational change and teachers' professional identities. *Teaching and Teacher Education, 47*, 1-12. doi:10.1016/j.tate.2014.11.006

van Ackeren, I. (2005). Vom Daten- zum Informationsreichtum? Erfahrungen mit standardisierten Vergleichstests in ausgewählten Nachbarländern. *Pädagogik, 57*(5), 24-28.

van Ackeren, I. (2007). *Nutzung großflächiger Tests für die Schulentwicklung. Exemplarische Analyse der Erfahrungen aus England, Frankreich und den Niederlanden*. Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).

van Ackeren, I., & Bellenberg, G. (2004). Parallelarbeiten, Vergleichsarbeiten und Zentrale Abschlussprüfungen. Bestandsaufnahme und Perspektiven. In H. G. Holtappels, K. Klemm, H. Pfeiffer, H.-G. Rolff, & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven. Band 13* (S. 125-159). Weinheim & München: Juventa.

van Ackeren, I., Block, R., Klein, E. D., & Kühn, S. M. (2012). The Impact of State-Wide Exit Exams in Germany: A Descriptive Case Study of Three German States with Differing Low Stakes Exam Regimes. *Education Policy Analysis Archives, 20*, 1-28.

van Ackeren, I., Brauckmann, S., & Klein, E. D. (2016). Internationale Diskussions-, Forschungs- und Theorieansätze zur Governance im Schulwesen. In H. Altrichter, & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (2., überarbeitete und aktualisierte Auflage, S. 29-51). Wiesbaden: Springer VS.

van Ackeren, I., & Klemm, K. (2011). *Entstehung, Struktur und Steuerung des deutschen Schulsystems. Eine Einführung* (2., aktualiserte und überarbeitete Auflage). Wiesbaden: Springer VS.

van Ackeren, I., Klemm, K., & Kühn, S. M. (2015). *Entstehung, Struktur und Steuerung des deutschen Schulsystems. Eine Einführung* (3., überarbeitete & aktualisierte Auflage). Wiesbaden: Springer VS. doi: 10.1007/978-3-531-20000-2

van Ackeren, I., Zlatkin-Troitschanskaia, O., Binnewies, C., Clausen, M., Dormann, C., Preisendörfer, P., Rosenbusch, C., & Schmidt, U. (2011). Evidenzbasierte Schulentwicklung. Ein Forschungsüberblick aus interdisziplinärer Perspektive. *Die deutsche Schule, 103*(2), 170-184.

van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of education review, 30*(5), 1045-1058. doi:10.1016/j.econedurev.2011.05.008

van Veen, K., Sleegers, P., & van de Ven, P.-H. (2005). One teacher's identity, emotions, and commitment to change: A case study into the cognitive-affective processes of a secondary school teacher in the context of reforms. *Teaching and Teacher Education, 21*, 917-934. doi:10.1016/j.tate.2005.06.004

Vanlaar, G., Kyriakides, L., Panayiotou, A., Vandecandelaere, M., McMahon, L., De Fraine, B., & Van Damme, J. (2016). Do the teacher and school factors of the dynamic model affect high- and low-achieving student groups to the same extent? a cross-country study. *Research Papers in Education, 31*(2), 183-211. doi:10.1080/02671522.2015.1027724

vbw – Vereinigung der Bayerischen Wirtschaft e. V. (2011). *Gemeinsames Kernabitur zur Sicherung von nationalen Bildungsstandards und fairem Hochschulzugang. Gutachten*. Münster: Waxmann.

von Recum, H. (2003). Aspekte bildungspolitischer Steuerung. In H. Döbert, B. von Kopp, R. Martini, & M. Weiß (Hrsg.), *Bildung vor neuen Herausforderungen. Historische Bezüge – Rechtliche Aspekte – Steuerungsfragen – Internationale Perspektiven* (S. 102-110). Neuwied: Luchterhand.

Watanabe, Y. (2004). Methodology in Washback Studies. In L. Cheng, Y. Watanabe, & A. E. Curtis (Hrsg.), *Washback in language testing. Research contexts and methods* (S. 19-36). Mahwah, NJ: Erlbaum.

Weick, K. E. (1976). Educational Organizations as Loosely Coupled Systems. *Administrative Science Quarterly, 21*, 1-19.

Weinstein, R. S. (2002). *Reaching higher. The power of expectations in schooling*. Cambridge, MA & London: Harvard University Press.

Wikström, C. (2005). Grade stability in a criterion-referenced grading system: the Swedish example. *Assessment in Education: Principles, Policy & Practice, 12*(2), 125-144. doi:10.1080/09695940500143811

Wiliam, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist, 45*(2), 107-122. doi:10.1080/00461521003703060

Wissenschaftsrat. (2012). *Prüfungsnoten an Hochschulen im Prüfungsjahr 2010 – Arbeitsbericht mit einem wissenschaftspolitischen Kommentar des Wissenschaftsrates*. Hamburg: Geschäftsstelle des Wissenschaftsrats.

Wissinger, J. (2007). Does School Governance matter? Herleitungen und Thesen aus dem Bereich „School Effectiveness and School Improvement". In H. Altrichter, T. Brüsemeister, & J. Wissinger (Hrsg.), *Educational Governance. Handlungskoordination und Steuerung im Bildungssystem* (S. 105-129). Wiesbaden: VS.

Woessmann, L., Luedemann, E., Schuetz, G., & West, M. R. (2009). *School Accountability, Autonomy and Choice around the World*. Cheltenham & Northhampton, MA: Edward Elgar.

Wößmann, L. (2003). Zentrale Prüfungen als „Währung" des Bildungssystems: Zur Komplementarität von Schulautonomie und Zentralprüfungen. *Vierteljahrshefte zur Wirtschaftsforschung, 72*(2), 220-237.

Wößmann, L., Lergetporer, P., Grewenig, E., Kugler, F., & Werner, K. (2017). Fürchten sich die Deutschen vor der Digitalisierung? Ergebnisse des ifo Bildungsbarometers 2017. *ifo Schnelldienst, 70*(17), 17-38.

Yeh, S. S. (2005). Limiting the Unintended Consequences of High-Stakes Testing. *Education Policy Analysis Archives, 13*(43), 1-21.

Zeidner, M. (1998). *Test anxiety: the state of the art*. New York: Plenum Press.

Zeidner, M., & Schleyer, E. J. (1998). The Big-Fish-Little-Pond Effect for Academic Self-Concept, Test Anxiety, and School Grades in Gifted Children. *Contemporary educational psychology, 24*(4), 305-329.

Zeitler, S. (2012). Forschungsstand. In S. Zeitler, N. Heller, & B. Asbrand (Hrsg.), *Bildungsstandards in der Schule. Eine rekonstruktive Studie zur Implementation der Bildungsstandards* (S. 23-47). Münster et al.: Waxmann.

Ziegelbauer, S. (2015). Akzeptanz als Voraussetzung gelingender Innovation in Schule. In J. Berkemeyer, N. Berkemeyer, & F. Meetz (Hrsg.), *Professionalisierung und Schulleitungshandeln. Wege und Strategien der Personalentwicklung an Schulen* (S. 146-159). Weinheim & Basel: Beltz Juventa.

Zlatkin-Troitschanskaia, O. (2007). Steuerungs*fähigkeit* des öffentlichen Schulwesens versus Steuer*barkeit* der Schule – Paradigmenwechsel? In J. van Buer, & C. Wagner (Hrsg.), *Qualität von Schule. Ein kritisches Handbuch* (S. 67-81). Frankfurt am Main et al.: Peter Lang.

Zlatkin-Troitschanskaia, O., Förster, M., & Preuße, D. (2012). Implementierung und Wirksamkeit der erweiterten Autonomie im öffentlichen Schulwesen – Eine Mehrebenenbetrachtung. In A. Wacker, U. Maier, & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (S. 79-107). Wiesbaden: Springer VS.

ANHANG

Publikation 1: Vergleichbarkeit der Abiturnoten in Mathematik

Maué, E. (2013). Vergleichbarkeit von Abiturnoten – eine Fiktion? Längerfristige Effekte der Implementation zentraler Abiturprüfungen in Bremen. In J. Asdonk, S. U. Kuhnen & P. Bornkessel (Hrsg.). Von der Schule zur Hochschule. Analysen, Konzeptionen und Gestaltungsperspektiven des Übergangs (S. 114-128). Münster: Waxmann.

Nicht, was Schüler lernen, bestimmt
ihren Schulerfolg, ihre Lebenschancen,
sondern wie sie zensiert werden.
(Ingenkamp, 1972)

1. EINLEITUNG

Schulabschlüsse stellen entscheidende „Stellschrauben" für den weiteren Lebensweg dar. Das gilt insbesondere für das Abitur – die formale Berechtigung für den Übertritt an eine Universität und damit für die Aufnahme eines Studiums. Obwohl sich die Zulassungsbedingungen im Wandel befinden und mittlerweile neben der Abiturdurchschnittsnote auch andere Faktoren an Bedeutung gewinnen (z.B. Sperlich, 2009), kommt ihr immer noch ein zentraler Stellenwert zu. Um Ungerechtigkeiten weitgehend zu vermeiden, sollten (Abitur-)Noten daher möglichst vergleichbar sein. Allerdings stehen bereits seit den 1970er Jahren die Objektivität, Reliabilität, Validität und Vergleichbarkeit von Noten sowie schriftlichen Arbeiten in der Kritik (Ingenkamp, 1972; Lintorf, 2012). Unbestritten ist, dass Noten nicht ausschließlich auf der zu beurteilenden Leistung beruhen, sondern darüber hinaus leistungsfremde Faktoren einfließen. Deshalb bezeichnet Ingenkamp (2005, S. 147) es als „Fiktion", dass „Zensuren über verschiedene Klassen vergleichbar wären". Um die Vergleichbarkeit von Leistungen und (Abschluss) Prüfungen zu sichern, wurden diverse Anstrengungen unternommen, wie etwa die Implementation von Vergleichsarbeiten und Bildungsstandards (z.B. Oelkers & Reusser, 2008), die Einführung und Modifikation von Einheitlichen Prüfungsanforderungen für die Abiturprüfung (EPA; vgl. Kultusministerkonferenz,

2008) oder zentral gestellte Abschlussprüfungen und Korrekturkriterien, welche selbst bei dezentraler Korrektur die Vergleichbarkeit von Schulabschlüssen gewährleisten sollen. Dabei ist im Fall von zentralen Abiturprüfungen jedoch kritisch anzumerken, dass erstens in einigen Bundesländern mit Zentralabitur (so auch in Bremen) weiterhin die Kursleitungen die Erstkorrektur und zumeist schulinterne Lehrkräfte die Zweitkorrektur übernehmen und zweitens die Benotung oftmals nicht anonymisiert erfolgt (Kühn, 2012). Dadurch besteht zwar Spielraum für die Berücksichtigung spezifischer Bedingungen, aber auch für Ungerechtigkeiten, z.B. durch Vorwissen und Vorannahmen über die Prüflinge, implizite Persönlichkeits- und Begabungstheorien der Lehrpersonen oder klasseninterne Bezugssysteme bei der Bewertung (Ingenkamp, 2005; Ditton, 2007a), was sich u.a. in „erhebliche[n] Maßstabsdifferenzen von Klasse zu Klasse und von Fach zu Fach" (Ingenkamp, 2005, S. 154) widerspiegelt.

Mit zentral gestellten Prüfungen ist darüber hinaus die Intention verbunden, die Vergleichbarkeit dezentral vergebener Noten zu steigern, indem sie eine Orientierung der Benotung an einem schulübergreifenden Kriterium bewirken (Standardisierung), sodass die Notengebung stärker auf der zu bewertenden Leistung basiert und weniger von leistungsfremden Merkmalen – wie Geschlecht, familiärer Bildungs- oder Migrationshintergrund – beeinflusst ist. Neumann et al. (2009) sowie Neumann, Trautwein und Nagy (2011) liefern erste empirische Hinweise, dass im Fach Mathematik „landesspezifische Zentralprüfungen [...] eine Annäherung länderübergreifender Bewertungsmaßstäbe" (Neumann et al., 2009, S. 707) und damit eine verbesserte Vergleichbarkeit der Abiturnoten initiieren. Da dieser Befund auf Querschnittanalysen beruht, ist allerdings unklar, inwiefern der Wechsel eines Prüfungssystems im Längsschnitt ebenfalls einen Beitrag zur Standardisierung der Beurteilung leistet. An diesem Punkt setzt der vorliegende Beitrag an, indem er prüft, ob die Implementation des Zentralabiturs in Bremer Mathematik-Leistungskursen mit einer kurz- und/oder längerfristigen Erhöhung der Vergleichbarkeit der Punktzahl im schriftlichen Mathematik-Abitur auf Bundeslandebene einhergeht.

2. FORSCHUNGSSTAND

Dass die Benotung einer Leistung zwischen Bundesländern, Schulen, Klassen, Kursen und Fächern unter anderem aufgrund differenzieller Lern- und Entwicklungsmilieus variiert, ist hinreichend belegt (Baumert & Watermann, 2000; Ingenkamp, 1972; Hochweber, 2010; Neumann et al., 2009; Schuler, 2006). Gleiches gilt für den Einfluss des individuellen und familiären Hintergrundes sowohl auf die Schulleistung als auch auf die Noten (Bornkessel & Kuhnen, 2011; Büchel, Jürges & Schneider, 2003; Maaz, Baeriswyl & Trautwein, 2011). Schülerinnen erreichen beispielsweise geringere Leistungen in Mathematik als Schüler (Mullis & Stemler, 2002), dennoch erhalten erstere bessere (Abitur-)Noten als ihre Mitschüler (Bornkessel & Kuhnen, 2011; Hochweber, 2010). Bezüglich des Bildungserfolgs von Kindern und Jugendlichen mit Migrationshintergrund und Benachteiligungen ihnen gegenüber ist die Forschungslage uneinheitlich und reicht von Benachteiligung bis Bevorzugung (z.B. Radtke, 2004; Stanat & Edele, 2011). Gresch (2012) sowie Asdonk und Sterzik (2011, S. 236f.) halten jedoch fest, dass „nicht der Migrationshintergrund an sich für den niedrigeren Abiturnotendurchschnitt verantwortlich ist, sondern die Tatsache, dass sich Schülerinnen und Schüler mit Migrationshintergrund bezüglich Bildung und sozioökonomischem Status strukturell unterscheiden". Einen weiteren Einfluss auf Noten und Übertrittsempfehlungen am Ende der Grundschulzeit üben Effekte der Referenzgruppe aus. Basierend auf dem von Marsh (1987) für das Selbstkonzept beschriebenen „Big-Fish-Little-Pond-Effekt" wiesen u.a. Trautwein und Baeriswyl (2007) einen Kompositionseffekt der Klasse auf die Übertrittsempfehlungen und die realisierten Übertritte nach.

Die Auswirkungen der Implementation des Zentralabiturs in Bremen und Hessen auf die Vergleichbarkeit der schriftlichen Abiturpunktzahl in Englisch- und Mathematik-Leistungskursen im Zeitraum 2007 bis 2009 untersuchte Holmeier (2012). Die Ergebnisse zeigen für die Mathematik-Leistungskurse, dass in beiden Bundesländern ein enger Zusammenhang zwischen der Punktzahl im schriftlichen Mathematik-Abitur und der individuellen Leistung im Mathematiktest (Kurztest aus TIMSS) bestand, wohingegen die mittlere Leistung des Kurses im Mathematiktest keinen Effekt aufwies. In Hessen fand in diesem Zeitraum für den Mathematik-Leistungskurs bei der Vergabe der Abiturpunktzahl keine Benachteiligung aufgrund des Geschlechts sowie des familiären Bildungshintergrundes (Bücheranzahl im Elternhaus als Indikator) statt. Jedoch erhielten im Ausland geborene Schülerinnen und Schüler bei gleicher Leistung im Mathematiktest tendenziell weniger Abiturpunkte.

Für die Mathematik-Leistungskurse in Bremen ist festzuhalten, dass Schüler bei gleichen Leistungen im Mathematiktest durchgängig eine geringere Punktzahl im schriftlichen Abitur in Mathematik erreichten als Schülerinnen, die Einführung des Zentralabiturs demnach keine Veränderung des geschlechtsspezifischen Benachteiligungseffekts initiierte. Hingegen wirkte sie sich positiv auf den ungünstigen Effekt eines Migrationshintergrundes aus, der sich im dezentralen Abitur zeigte, 2008 und 2009 aber nicht mehr von Bedeutung war. Kurzzeitig führte die Implementation des Zentralabiturs zudem zu einer Benachteiligung von Jugendlichen aus bildungsferneren Elternhäusern, was jedoch 2009 nicht mehr der Fall war. Insgesamt erhöhte das Zentralabitur in Bremer Mathematik-Leistungskursen zwar nicht den Zusammenhang zwischen der Leistung im Mathematiktest und im schriftlichen Abitur, allerdings verringerte sich der Einfluss des Migrationshintergrundes bis zum Jahr 2009. Während sich der mit der Einführung des Zentralabiturs einhergehende negative Effekt des familiären Bildungshintergrunds bis 2009 wieder reduzierte, blieb der Effekt des Geschlechts über die Jahre bestehen.

In Ergänzung zu den Resultaten von Holmeier (2012) nimmt der vorliegende Beitrag längerfristige Entwicklungen bis zum Jahr 2011 in den Blick. Dabei stehen eine mögliche parallele Entwicklung der Mathematikleistungen und der Abiturpunktzahl sowie potentielle Standardisierungseffekte (stärkere Orientierung der Bewertung an einem externen, schulübergreifenden Kriterium, geringerer Einfluss leistungsfremder Faktoren), die zu einer besseren Vergleichbarkeit der Abiturnoten beitragen, im Zentrum des Interesses.

3. FRAGESTELLUNG UND HYPOTHESEN

Der vorliegende Beitrag richtet seinen Fokus auf folgende Fragestellungen:

Inwiefern zeigen sich in Bremen parallele Entwicklungen für die im Mathematiktest erzielte Leistung und für die im schriftlichen Abitur im Mathematik-Leistungskurs erreichte Punktzahl im Zeitraum von 2007 bis 2011?
Holmeier (2012) wies für den Leistungskurs im Fach Mathematik in Bremen für den Zeitraum von 2007 bis 2009 nach, dass die Veränderungen in den Mathematikleistungen (niedrigste Werte in 2008) nicht mit den Veränderungen in der Abiturpunktzahl (niedrigste Werte in 2007) übereinstimmten. Da sich

Effekte der Implementation des Zentralabiturs erst mit der Zeit zeigen können, wird in längerfristiger Perspektive jedoch eine Angleichung der Entwicklungen erwartet (Hypothese 1).

Wie gestaltet sich in den Jahren 2007 bis 2011 der Zusammenhang von der Leistung im Mathematiktest und im schriftlichen Abitur? Zeigen sich prüfungsformspezifische Effekte?
Es ist anzunehmen, dass die Implementation des Zentralabiturs mit einer stärkeren Orientierung an einem externen, schulübergreifenden Maßstab einhergeht und dies in einem über die Jahre engeren Zusammenhang zwischen der Leistung im Mathematiktest und der Abiturpunktzahl resultiert (Hypothese 2). Zwar konnte Holmeier (2012) für den Zeitraum 2007 bis 2009 keine Veränderungen feststellen, da eine standardisierende Wirkung auf die Beurteilung Zeit und Erfahrung benötigt (Maag Merki, 2012a), sollte sich jedoch zumindest in längerfristiger Perspektive ein Effekt zeigen.

Welchen Beitrag leisten zentrale Abiturprüfungen zur Kompensation der Einflüsse leistungsfremder Faktoren? Findet durch das Zentralabitur eine Verringerung bzw. Verhinderung von Benachteiligungen statt?
Zunächst ist davon auszugehen, dass in Übereinstimmung mit den Befunden von Holmeier (2012), Maaz et al. (2011) und Neumann et al. (2009; 2011) die Abiturpunktzahl in starkem Maß von den Leistungen im Mathematiktest auf Individualebene bestimmt wird (Hypothese 3). Zudem sollte sich aufgrund von Standardisierungseffekten der Einfluss der Mathematikleistungen auf die Bewertung der Abiturleistung in den vier Jahren seit der Implementation vergrößert haben (Hypothese 4).

Dennoch dürften leistungsfremde Faktoren nach wie vor eine Rolle spielen (z.B. Bornkessel & Kuhnen, 2011; Maaz et al., 2011). Innerhalb des Zeitraumes 2007 bis 2009 beeinflusste von den leistungsfremden Merkmalen vorrangig das Geschlecht der jungen Erwachsenen die Abiturpunktzahl. Der kurzfristig mit der Einführung des Zentralabiturs einhergehende ungünstige Effekt einer geringen Bücherzahl im Elternhaus im Jahr 2008 ist 2009 wieder zurückgegangen. Im Gegensatz zum dezentralen Prüfsystem erwies sich im zentralen ein ausländisches Geburtsland als weniger (2008) bzw. nicht mehr nachteilig (2009; Holmeier, 2012). Darauf basierend wird angenommen, dass sich diese Entwicklungen fortsetzen und die leistungsfremden Faktoren Geschlecht, familiärer Bildungshintergrund (Anzahl der Bücher im Elternhaus als Indikator) und Geburtsland mit der Zeit einen geringeren bzw. keinen Einfluss (mehr) aufweisen (Hypothese 5).

4. STUDIE „IMPLEMENTATION UND AUSWIRKUNGEN NEUER STEUERUNGSSTRUKTUREN IM SCHULWESEN AM BEISPIEL ZENTRALER ABITURPRÜFUNGEN“

Die dem Beitrag zu Grunde liegende Studie wurde an der Universität Zürich in Kooperation mit dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF) in den Jahren 2007 bis 2009 sowie im Jahr 2011 durchgeführt. In Bremen erfolgte das Zentralabitur 2011 in den Grundkursen zum fünften, in den Leistungskursen in einzelnen Fächern (Mathematik, Deutsch, fortgesetzte Fremdsprache, Naturwissenschaften) zum vierten Mal, was die Untersuchung längerfristiger Effekte der Einführung ermöglicht.

Design

Ziel der Studie ist die Analyse der „Implementation zentral organisierter Abiturprüfungen als ein Element im neuen Konzept der Systemsteuerung in den zwei deutschen Bundesländern Bremen und Hessen [...]. Im Zentrum stehen Fragen a) zu den Effekten des *Wechsels* [Hervorhebung im Original; E. M.] von einem dezentralen zu einem zentralen Prüfungssystem in Bremen für Schüler/-innen, Lehrpersonen, Unterricht und Schule sowie b) zu den Veränderungen des schulischen Handelns und der schulischen Leistungen *nach Implementation zentraler Abiturprüfungen* [Hervorhebung im Original; E. M.] in beiden Bundesländern“ (Maag Merki, 2012a, S. 13). Hierfür wurden in den Jahren 2007, 2008, 2009 und 2011 Erhebungen bei Schülerinnen und Schülern, Lehrpersonen und Schulleitungen (nur 2011) durchgeführt. Einen Überblick über das Design und die eingesetzten Instrumente der Jahre 2007 bis 2009 bieten Maag Merki und Oerke (2012). Für den vorliegenden Beitrag sind der Mathematikleistungstest, die Abiturpunktzahl sowie ausgewählte Angaben zum individuellen und familiären Hintergrund der Bremer Schülerinnen und Schüler der Mathematik-Leistungskurse von Interesse (vgl. Tabelle 1).

Tabelle 1: In den Analysen berücksichtigte Variablen

Variable	Ausprägung
Abiturpunktzahl	0 – 15 Punkte
Mathematikleistungstest	0 – 15 Punkte
Geschlecht	0 = weiblich; 1 = männlich
Geburtsland	0 = Deutschland; 1 = Ausland
Buchbestand im Elternhaus	0 = 0 bis 10; 1 = 11 bis 50; 2 = 51 bis 100; 3 = 101 bis 250; 4 = 251 bis 500; 5 = mehr als 500

Quelle: Eigene Darstellung.

Mittels 15 Items des TIMSS-Tests „Fachleistungen im voruniversitären Mathematikunterricht" (Klieme, 2000) werden die Mathematikleistungen der Abiturientinnen und Abiturienten erfasst. Dieser Test hat sich als curricular valide erwiesen (Klieme, 2000), zudem liegt die Reliabilität im akzeptablen bis guten Bereich (Reliabilität: 2007: $\alpha = .80$, N = 242; 2008: $\alpha = .67$, N = 317; 2011: $\alpha = .77$, N = 612). Für die folgenden Analysen wird der Summenscore der 15 Items herangezogen.

Methodik/Auswertungsstrategien

Zur Nachzeichnung der Entwicklung in Bremer Mathematik-Leistungskursen gehen in die Berechnungen die Daten der Jahre 2007, 2008 und 2011 ein. Detaillierte Auswertungen der Jahre 2007 bis 2009 bieten Holmeier (2012) und Maag Merki (2012b). Einen ersten Überblick über die Resultate geben die deskriptive Verteilung der Punktzahl im schriftlichen Abitur und im Leistungstest sowie deren Korrelationen (Prüfung auf signifikante Differenzen zwischen den Jahren mittels Fishers Z-Transformation). Varianzanalysen dienen der Aufdeckung möglicher Jahresunterschiede bezüglich der Abitur- und der Leistungstestpunktzahl. Mehrebenenanalysen sollen die Einflüsse verschiedener Faktoren auf Individual- und Kursebene auf die Punktzahl im schriftlichen Mathematik-Abitur erklären. Alle Variablen gehen unzentriert in die Analysen ein. Lediglich der Mathematikleistungstest wird auf Level 1 und Level 2 zentriert (grand mean), sodass das Gesamtmodell (Intercepts-andSlopes-as-Outcomes-Modell) in Analogie zu Holmeier (2012) auf folgender Gleichung beruht (Berechnung mit HLM 6.06; Raudenbush, Bryk & Congdon, 2004):

*Punktzahl im schriftlichen Abitur = y00 + y01*Mathetest_Level2 + y02*Jahr2008 + 03*Jahr2011 + y10*Geschlecht + y11*(Jahr2008*Geschlecht) + y12*(Jahr2011*Geschlecht) + y20*Geburtsland + y21*(Jahr2008*Geburtsland) + y22*(Jahr2011*Geburtsland) + y30*Bücheranzahl + y31 *(Jahr2008*Bücheranzahl) + y32*(Jahr2011*Bücheranzahl) + y40*Mathetest_Level1 + y41 *(Jahr2008*Mathetest_Level1) + y42*(Jahr2011*Mathetest_Level1) + u0 + u1*Geschlecht + u2*Geburtsland + u3*Bücheranzahl + u4*Mathetest_Level1 + r*

Stichprobe

Die Berechnungen stützen sich auf die Daten der Schülerinnen und Schüler in Bremen der Jahre 2007, 2008 und 2011. Dabei handelt es sich auf individueller Ebene um querschnittliche Analysen, da die Abiturientinnen und Abiturienten jeweils nur einmal befragt wurden. Jedes Jahr nahm je ein Mathematik-Leistungskurs pro Schule an den Erhebungen teil. Da über die Jahre immer die gleichen Schulen teilnahmen, ergibt sich auf Schulebene ein Längsschnitt. Es finden ausschließlich die Schulen Berücksichtigung, in deren Mathematik-Leistungskursen für jedes Jahr die Daten von mindestens fünf Abiturientinnen und Abiturienten vorliegen. Insgesamt kann auf 33 Kurse aus 11 Schulen zurückgegriffen werden.

Während der Anteil der im Ausland geborenen Abiturientinnen und Abiturienten über die Jahre relativ stabil blieb, fanden sich im Jahr 2008 deutlich mehr junge Frauen in der Stichprobe. Dies ist bei den folgenden Analysen zu bedenken.

Tabelle 2: Stichprobe der Schülerinnen und Schüler in Bremen nach Jahren

	Kohorten		
Variable	**2007**	**2008**	**2011**
Abiturpunktzahl	158	153	192
Leistungstest	152	158	195
Geschlecht (davon weiblich)	158 (32%)	157 (42%)	196 (30%)
Geburtsland (davon Ausland)	126 (12%)	139 (15%)	183 (13%)
Buchbestand im Elternhaus	128	139	180

Quelle: Eigene Berechnungen.

5. ERGEBNISSE

Zur Beantwortung der Frage, ob die Entwicklung der Punktzahl im schriftlichen Mathematik-Abitur und im Mathematikleistungstest parallel verläuft (Fragestellung 1), werden die einzelnen Jahre 2007, 2008 und 2011 deskriptiv ausgewertet.

Tabelle 3: Deskriptive Statistik der Punktzahl im schriftlichen Abitur und im Mathematik-Leistungstest nach Jahren

	Abiturpunktzahl			Leistungstest			
Jahr	**N**	**M**	**SD**	**N**	**M**	**SD**	**r**
2007	158	8.80	3.69	152	8.47	3.38	.41***
2008	153	9.45	3.52	158	7.89	2.66	.48***
2011	192	8.53	3.70	195	8.78	2.75	.48***

Anmerkungen: N = Stichprobengröße; *M* = Mittelwert; *SD* = Standardabweichung; *r* = Produkt-Moment-Korrelation.
Quelle: Eigene Berechnungen.

Die Schülerinnen und Schüler erhielten im Jahr 2008 tendenziell mehr Punkte im schriftlichen Abitur als im Jahr 2011 ($d = .25^{+}$), obwohl sie in 2008 signifikant weniger Punkte im Mathematikleistungstest erzielten als in 2011 ($d = -.33^{*}$). Abgesehen von diesen beiden Effekten zwischen den Jahren 2008 und 2011 liegen weder für die Abiturpunktzahl noch für den Leistungstest in Mathematik signifikante Differenzen zwischen den Jahren 2007 und 2008 sowie zwischen 2007 und 2011 vor. Dies belegen Varianzanalysen, die zwischen diesen Jahren keine signifikanten Differenzen anzeigen (Abiturpunktzahl: 2007-2008: $d = -.18$, n.s.; 2007-2011: $d = .07$, n.s.; Mathematiktest: 2007-2008: $d = .19$, n.s.; 2007-2011: $d = -.10$, n.s.). Damit ergeben sich in der 5-Jahres-Perspektive keine bedeutsamen Veränderungen im Leistungsniveau und in der Bewertung der Schülerinnen und Schüler.

Ob sich der Zusammenhang zwischen der Leistung im Mathematiktest und im schriftlichen Mathematik-Abitur mit der Zeit verstärkt (Fragestellung 2), wird mittels Korrelationen überprüft. In allen drei untersuchten Jahren fallen diese im Mittel in einer ähnlichen Höhe aus ($r = .41^{***}$ bis $r = .48^{***}$). Signifikante Differenzen zwischen den Korrelationen der verschiedenen Jahre lassen sich nicht ausmachen, der Zusammenhang wird demzufolge mit der Zeit nicht enger und gestaltet sich nicht in Abhängigkeit

der Prüfungsform. Die Mehrebenenanalysen zur Überprüfung des Effekts leistungsfremder Faktoren auf die Benotung des schriftlichen Mathematik-Abiturs wurden zunächst für jeden Einflussfaktor separat berechnet und anschließend in einem Gesamtmodell integriert (Fragestellung 3, Tabelle 4). Dabei zeigen sich lediglich geringe Abweichungen zwischen den verschiedenen Modellen, beispielsweise schwanken die Effekte von Geburtsland und Anzahl der Bücher im Elternhaus um die Grenze von signifikantem und tendenziell signifikantem Niveau. Aus Gründen der Übersichtlichkeit wird hier lediglich das Gesamtmodell präsentiert.

Auf individueller Ebene stehen die Mathematikleistungen in allen drei Jahren in engem Zusammenhang mit dem Abschneiden im schriftlichen Mathematik-Abitur (höchst signifikanter Haupteffekt, Interaktionseffekte nicht signifikant). Das Geburtsland spielt ebenfalls eine Rolle: Im dezentralen Abitur (2007) werden junge Erwachsene mit Migrationshintergrund bei gleicher Leistung schlechter bewertet (signifikanter Haupteffekt). Dieser Effekt schwächt sich mit der Implementation des Zentralabiturs im Jahr 2008 ab (signifikanter Interaktionseffekt), ist allerdings vier Jahre später im Jahr 2011 wieder existent (Interaktionseffekt nicht signifikant). Während bei dezentraler Prüfungsorganisation kein Effekt des familiären Bildungshintergrundes auszumachen ist (Haupteffekt nicht signifikant), tritt dieser mit Einführung des Zentralabiturs tendenziell auf (tendenziell signifikanter Interaktionseffekt): Abiturientinnen und Abiturienten aus bildungsferneren Familien erhalten bei gleichen Leistungen im Jahr 2008 weniger Punkte im schriftlichen Mathematik-Abitur et vice versa. Allerdings scheint es sich wie bei Holmeier (2012) lediglich um einen kurzfristigen Effekt zu handeln, da 2011 kein Unterschied mehr zu beobachten ist (kein signifikanter Interaktionseffekt). In allen drei Jahren sind geschlechtsspezifische Differenzen nicht von Bedeutung (Haupteffekt und Interaktionseffekte nicht signifikant). Unter Kontrolle der Leistung werden demnach Schülerinnen und Schüler im schriftlichen Mathematik-Abitur nicht unterschiedlich beurteilt. Insgesamt vermag die Berücksichtigung der Struktur des Mehrebenensystems Schule (Ditton, 2007b; Fend, 2008) 14 Prozent der Varianz bezüglich der Punktzahl im schriftlichen Abitur im Leistungskurs Mathematik zu erklären. Weiteres Aufklärungspotential besteht bezüglich der Varianz des Geschlechtereffektes (signifikante Varianzkomponente).

Tabelle 4: Mehrebenenanalyse: Effekte der individuellen und aggregierten Mathematikleistung, des Geschlechts, Geburtslandes und der Bücheranzahl im Elternhaus sowie der Zeit auf die Punktzahl im schriftlichen Mathematik-Abitur

Fixe Effekte	**nicht stand. Koeffizienten (robuste Standardfehler)**
Ebene 1	
Konstante	9.48 (1.31) ***
Leistungstest	0.60 (0.11) ***
Geschlecht (0 = weiblich, 1 = männlich)	-0.70 (0.57) n.s.
Geburtsland (0 = Deutschland, 1 = Ausland)	-2.01 (0.94) *
Bücheranzahl	0.06 (0.27) n.s.
Ebene 2	
Leistungstest	-0.14 (0.10) n.s.
Jahr2008 (2008 = 1)[1]	-0.97 (1.36) n.s.
Jahr2011 (2011 = 1)[2]	-2.16 (1.43) n.s.
Interaktionseffekte	
Jahr2008*Leistungstest	-0.15 (0.15) n.s.
Jahr2008*Geschlecht	-0.54 (0.80) n.s.
Jahr2008*Geburtsland	2.20 (1.03) *
Jahr2008*Bücheranzahl	0.55 (0.28) +
Jahr2011*Leistungstest	0.05 (0.13) n.s.
Jahr2011*Geschlecht	0.97 (0.84) n.s.
Jahr2011*Geburtsland	1.71 (1.04) n.s.
Jahr2011*Bücheranzahl	0.25 (0.29) n.s.
Zufällige Effekte	
Varianzkomponenten	
u_0	1.23 (1.11) n.s.
u_1 (Geschlecht)	0.89 (0.94) *
u_2 (Geburtsland)	0.25 (0.50) n.s.
u_3 (Bücheranzahl)	0.01 (0.10) n.s.
u_4 (Leistungstest)	0.01 (0.09) n.s.
R	8.87 (2.98)
Intraclass-Correlation	*0.14*

Anmerkungen: [1] 2008 im Vergleich zu 2007: negativer Wert: Abnahme von 2007 zu 2008
[2] 2011 im Vergleich zu 2007: negativer Wert: Abnahme von 2007 zu 2011

Quelle: Eigene Berechnungen.

Die Punktzahl im schriftlichen Mathematik-Abitur ändert sich über die Jahre nicht, da auf Level 2 die Effekte der Jahre 2008 und 2011 jeweils im Vergleich zu 2007 nicht signifikant ausfallen. Zudem sind keine Referenzgruppeneffekte im Sinne des Big-Fish-Litte-Pond-Effekts festzustellen, da der Einfluss des mittleren Leistungsniveaus des Kurses auf die Abiturbeurteilung ebenfalls nicht signifikant ist.

6. DISKUSSION

Die Nachzeichnung der Entwicklung der Punktzahl sowohl im Mathematiktest als auch im schriftlichen Abitur im Mathematik-Leistungskurs erfolgte zunächst über den Vergleich der Mittelwerte der einzelnen Jahre. Im Jahr der Einführung zentraler Abiturprüfungen im Leistungskurs Mathematik (2008) zeigten die Schülerinnen und Schüler signifikant niedrigere Leistungen als im Jahr 2011. Damit ist zwar eine Steigerung im Zeitraum 2008 bis 2011 zu verzeichnen, der 5-Jahres-Vergleich (2007-2011) verdeutlicht jedoch keine signifikante Zunahme der Leistungen und damit keinen gegen den Zufall abgesicherten Unterschied zwischen dezentralem und zentralem Abitur.

Bezüglich der im schriftlichen Mathematik-Abitur erreichten Punktzahl nimmt das Jahr 2008 ebenfalls eine Sonderposition ein, da dort, trotz der niedrigeren Leistungen im Mathematiktest, tendenziell eine höhere Punktzahl vergeben wurde als im Jahr 2011. Möglicherweise bewerteten in jenem Jahr die Lehrpersonen aufgrund der Einführung des Zentralabiturs etwas milder als in den beiden anderen Jahren oder es zeigen sich individuelle Bewertungsunterschiede zwischen den jeweils beteiligten Lehrpersonen. Da sich die Mathematikleistungen und die Abiturpunktzahl in der längerfristigen Perspektive nicht signifikant ändern und beim Jahr 2008 eine gegenläufige Entwicklung erfolgt, ist Hypothese 1 falsifiziert.

Die Höhe der Korrelationen zwischen Mathematiktest und Punktzahl im schriftlichen Mathematik-Abitur zwischen $r = .41$ *** und $r = .48$*** decken sich mit Befunden zum Zusammenhang von mathematischer Testleistung und Halbjahresnote von Hochweber (2010) und Levin (2009) für die Sekundarstufe I. Sie fallen hingegen etwas geringer aus als bei Neumann et al. (2009) für Hamburger und Baden-Württemberger Abiturientinnen und Abiturienten berichtet, ebenfalls bei Verwendung der Halbjahreszensuren

(vgl. für internationale Befunde Neumann et al., 2011). Entgegen der Erwartung (Hypothese 2) hat sich der Zusammenhang weder verstärkt noch übertrifft er die bei Holmeier (2012) berichteten Größenordnungen. Damit ist eine standardisierende Wirkung des Zentralabiturs aufgrund einer stärkeren Orientierung an einem externen, nicht schulspezifischen Kriterium nicht erkennbar. Ein Grund dafür könnte der Stellenwert subjektiver Persönlichkeits- und Begabungstheorien der Lehrpersonen bei der Bewertung sein (Ditton, 2007a).

Abschließend konnte die Berechnung von Mehrebenenanalysen Einflüsse auf die Punktzahl im schriftlichen Mathematik-Abitur in den Leistungskursen aufklären. Diese bleibt über die Zeit konstant und ändert sich nicht signifikant, sodass im dezentralen wie im zentralen Abitur ähnlich viele Punkte vergeben werden. Ein Effekt der mittleren Leistung auf Kursniveau ist nicht auszumachen, was im Gegensatz zu den Befunden der TOSCA-Studie, wo Hinweise auf einen Big-Fish-Little-Pond-Effekt vorliegen (Neumann et al., 2009), steht. In Übereinstimmung mit den Ergebnissen von Holmeier (2012) und Maaz et al. (2011) sowie mit Hypothese 3 spielt die individuelle Mathematikleistung der jungen Erwachsenen die größte Rolle bei der Bewertung der schriftlichen Abiturprüfung. Allerdings ist im Gegensatz zur vierten Hypothese keine Steigerung des Einflusses über die Zeit festzustellen, sodass sich die Notengebung im dezentralen wie im zentralen Prüfsystem im selben Ausmaß an den Mathematikleistungen orientiert. Offen bleiben muss an dieser Stelle, welche strukturellen, intra- und interindividuellen Einflüsse die Entstehung und Entwicklung letzterer mitbestimmt haben (z. B. Ditton, 2007a).

Neben fachlichen Fähigkeiten ist von Bedeutung, ob die Schülerinnen und Schüler im Ausland geboren wurden. Bei dezentraler Prüfungsorganisation erhielten sie bei gleicher Leistung signifikant weniger Punkte in der schriftlichen Abiturprüfung in Mathematik. Die Einführung zentraler Prüfungen trug zur Verringerung der Benachteiligung zugewanderter junger Erwachsener bei. Da diese Entwicklung jedoch nicht von längerfristiger Dauer war und sich der Migrationshintergrund drei Jahre später erneut ungünstig auf die Abiturbeurteilung auswirkte, könnte die geringere Benachteiligung im Jahr 2008 mit allgemein milderen Bewertungen aufgrund der Implementation zentraler Prüfungen im Leistungskurs oder mit anderen an der Korrektur beteiligten Lehrkräften begründet sein. Die Benachteiligung der im Ausland Geborenen im schriftlichen Mathematik-Abitur kann als eine Fortsetzung der „Bildungsnachteile[n] von Schülern mit Migrationshintergrund" (Diefenbach, 2009, S. 452) gesehen werden.

Es ist kritisch anzumerken, dass der Migrationshintergrund in der vorliegenden Studie lediglich mittels des Geburtslandes der Abiturientinnen und Abiturienten (Deutschland oder ein anderes Land) erfasst wurde und somit eine Differenzierung nach Herkunftsland und Generation nicht möglich ist. In Deutschland geborene Jugendliche mit deutschen oder ausländischen Eltern sind in diesem Fall nicht zu unterscheiden.

In Bremer Mathematik-Leistungskursen ist im Zeitraum 2007 bis 2011 keine geschlechtsspezifische Benachteiligung festzustellen – und zwar unabhängig vom Prüfsystem. Dies steht im Gegensatz zu den Befunden von Holmeier (2012) für Bremen (2007 bis 2009) und Maaz et al. (2011) für Baden-Württemberg, laut denen Abiturientinnen im Mathematik-Leistungskurs bei schwächeren Leistungen bessere Noten erhalten. Die Differenzen zu den Resultaten von Holmeier (2012) sind vermutlich in der Unterschiedlichkeit der Stichproben und des Referenzjahres begründet.

Ein durch zentrale Abiturprüfungen nicht intendierter Effekt zeigt sich in der tendenziellen Bevorzugung von Schülerinnen und Schülern mit einem vorteilhaften familiären Bildungshintergrund im Jahr 2008: Sie erhielten bei gleichen Leistungen tendenziell mehr Punkte im schriftlichen Mathematik-Abitur. Dieser Einfluss verringerte sich in längerfristiger Perspektive jedoch wieder, sodass im Jahr 2011 kein Unterschied zum dezentralen Abitur bestand. Dies entspricht sowohl den Befunden von Holmeier (2012) als auch der TOSCA-Studie (Maaz et al., 2011). Zu bedenken ist, dass in den vorliegenden Analysen einzig die Anzahl der Bücher im Elternhaus zur Operationalisierung des familiären Bildungshintergrundes herangezogen werden konnte und eine Differenzierung zwischen Struktur- und Prozessmerkmalen, wie beispielsweise bei Bornkessel und Kuhnen (2011), nicht möglich war.

Hypothese 5, die von einer Abnahme leistungsfremder Faktoren über die Zeit ausgeht, ist zu verwerfen. Einerseits existieren weder im dezentralen noch im zentralen Abitur Differenzen in Abhängigkeit vom Geschlecht und familiären Bildungshintergrund (zumindest im Vergleich 2007 bis 2011). Andererseits dauert die bereits 2007 bestehende Benachteiligung von Abiturientinnen und Abiturienten mit Migrationshintergrund im Jahr 2011 an. Gerade letztem Aspekt muss weitere Beachtung geschenkt und über Lösungsstrategien nachgedacht werden. Dies beinhaltet unter anderem eine Sensibilisierung der Lehrpersonen – und zwar nicht erst am Ende der Sekundarstufe II, sondern bereits in der Grundschule,

um u.a. der frühen Selektion von Kindern mit Migrationsgeschichte auf niedriger qualifizierende Sekundarschulen zu begegnen (Diefenbach, 2009).

Fazit

Es bleibt festzuhalten, dass sich die Implementation zentraler Abiturprüfungen in Bremer Mathematik-Leistungskursen in kurzfristiger Perspektive (2007 bis 2008) zwar positiv hinsichtlich der Benachteiligung von im Ausland geborenen Schülerinnen und Schülern auswirkte (Verringerung), jedoch mit negativen Folgen für Abiturientinnen und Abiturienten mit wenigen Büchern im Elternhaus einherging (tendenzielle Benachteiligung). Beide Gruppen erhielten bei gleicher Leistung mehr bzw. weniger Punkte im schriftlichen Mathematik-Abitur. In längerfristiger Perspektive (2007 bis 2011) kehrten sich diese Effekte um, sodass der „Ausgangszustand" und somit die Benachteiligungen von 2007 im dezentralen Abitur wiederhergestellt ist bzw. sind. Insgesamt zeichnen sich durch die Einführung des Zentralabiturs in Bremer Mathematik-Leistungskursen keine bedeutenden Veränderungen hin zu größerer Standardisierung und Vergleichbarkeit und damit Fairness der Abiturnote ab, da der hohe Einfluss der individuellen Mathematikleistungen sowie die nach Geschlecht und familiärem Bildungshintergrund gleiche Benotung auch im dezentralen Abitur Bestand hatte.

Kritisch anzumerken ist, dass die Analysen auf dem querschnittlichen Vergleich dreier Kohorten, die sich hinsichtlich der Zusammensetzung nach Geschlecht unterscheiden, basieren. Worauf die Differenzen zwischen den Jahren zurückzuführen sind und ob sich dadurch Einschränkungen für die Vergleichbarkeit der Befunde ergeben, müssen weitere Analysen klären. Darüber hinaus ist beim Vergleich mit der LAU- und TOSCA-Studie zu berücksichtigen, dass in Bremer Mathematik-Leistungskursen der Anteil der Abiturientinnen und Abiturienten mit Migrationshintergrund in etwa so groß wie unter Hamburger (Lehmann et al., 2012) jedoch annähernd doppelt so groß wie unter Baden-Württemberger Gymnasiastinnen und Gymnasiasten (Köller et al., 2004) ausfällt. Zudem konnte pro Jahr nur auf elf von maximal 19 Level 2-Einheiten zurückgegriffen werden, was die Repräsentativität einschränkt. In der Studie liegen zwar längsschnittliche Daten der Lehrpersonen vor, eine Zuordnung zu den jeweiligen Mathematik-Leistungskursen ist allerdings aufgrund der strikten Datenschutzbestimmungen nicht möglich. Dies böte eine ergänzende Perspektive auf potentielle Effekte der Implementation zentraler Abiturprüfungen auf die Benotung.

7. AUSBLICK

Dass Noten nach wie vor wenig vergleichbar sind, zeigte eindrücklich der Vergleich der TOSCA- und LAU-Studie (Trautwein et al., 2007). Auch die vorliegende Studie gibt, zumindest für Bremer Mathematik-Leistungskurse, keine Hinweise auf die Realisierung der zu Beginn als „Fiktion" zitierten Vergleichbarkeit der Noten durch die Implementation des Zentralabiturs. Insofern gilt es, den Fokus neben Maßnahmen zur Erhöhung der Vergleichbarkeit der Noten und Abschlüsse auch darauf zu richten, wie die Chancengleichheit verschiedener Gruppen im Bildungssystem verbessert werden kann, damit meritokratische Prinzipien anstelle von Zensuren Schulerfolg und Lebenschancen beeinflussen.

LITERATUR

Asdonk, J. & Sterzik, C. (2011). Kompetenzen für den Übergang zur Hochschule. In P. Bornkessel & J. Asdonk (Hrsg.), *Der Übergang Schule – Hochschule. Zur Bedeutung sozialer, persönlicher und institutioneller Faktoren am Ende der Sekundarstufe II* (S. 191–249). Wiesbaden: VS.

Baumert, J. & Watermann, R. (2000). Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), TIMSS/III. *Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 317–372). Opladen: Leske + Budrich.

Bornkessel, P. & Kuhnen, S. U. (2011). Zum Einfluss der sozialen Herkunft auf Schulleistung, Studienzuversicht und Studienintention am Ende der Sekundarstufe II. In P. Bornkessel & J. Asdonk (Hrsg.), *Der Übergang Schule – Hochschule. Zur Bedeutung sozialer, persönlicher und institutioneller Faktoren am Ende der Sekundarstufe II* (S. 47–104). Wiesbaden: VS.

Büchel, F., Jürges, H. & Schneider, K. (2003). Die Auswirkungen zentraler Abschlussprüfungen auf die Schulleistung – Quasi-experimentelle Befunde aus der deutschen TIMSS-Stichprobe. *Vierteljahreshefte zur Wirtschaftsforschung, 72* (2), 238–251.

Diefenbach, H. (2009). Der Bildungserfolg von Schülern mit Migrationshintergrund im Vergleich zu Schülern ohne Migrationshintergrund. In R. Becker (Hrsg.), *Lehrbuch der Bildungssoziologie* (S. 433–457). Wiesbaden: VS.

Ditton, H. (2007). Schulqualität – Modelle zwischen Konstruktion, empirischen Befunden und Implementierung. In J. van Buer & C. Wagner (Hrsg.), *Qualität von Schule. Ein kritisches Handbuch* (S. 83–92). Frankfurt am Main: Peter Lang.

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS.

Gresch, C. (2012). *Der Übergang in die Sekundarstufe I. Leistungsbeurteilung, Bildungsaspiration und rechtlicher Kontext bei Kindern mit Migrationshintergrund*. Wiesbaden: VS.

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf Schüler- und Schulklassenebene in Primar- und Sekundarstufe.* Münster: Waxmann.

Holmeier, M. (2012). Vergleichbarkeit der Punktzahlen im schriftlichen Abitur. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 289–320). Wiesbaden: VS.

Ingenkamp, K. (2005). *Lehrbuch der Pädagogischen Diagnostik* (5. völlig überarbeitete Aufl.). Weinheim: Beltz.

Ingenkamp, K. (Hrsg.). (1972). *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte* (3. Aufl.). Weinheim: Beltz.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 57–128). Opladen: Leske + Budrich.

Köller, O., Watermann, R., Trautwein, U. & Lüdtke, O. (Hrsg.). (2004). *Wege zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien.* Opladen: Leske + Budrich.

Kühn, S. M. (2012). Zentrale Abiturprüfungen im nationalen und internationalen Vergleich mit besonderer Perspektive auf Bremen und Hessen. In K. Maag Merki (Hrsg.), Zentralabitur. *Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 25–42). Wiesbaden: VS.

Kultusministerkonferenz (2008). *Vereinbarung über Einheitliche Prüfungsanforderungen in der Abiturprüfung.* Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2008/2008_10_24-VB-EPA.pdf [18.06.2012].

Levin, A. (2009). *Qualitätsprobleme mathematischer Vergleichsarbeiten. Erfassung mathematischer Kompetenzen und psychometrische Modellierung einer landesweiten Prüfungsarbeit in Klassenstufe 10*. Münster: Waxmann.

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Wiesbaden: VS.

Maag Merki, K. (2012a). Forschungsfragen und theoretisches Rahmenmodell. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 9–23). Wiesbaden: VS.

Maag Merki, K. (2012b). Die Leistungen der Gymnasiastinnen und Gymnasiasten in Mathematik und Englisch. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 259–288). Wiesbaden: VS.

Maag Merki, K. & Oerke, B. (2012). Methodische Grundlagen der Studie. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (S. 43–59). Wiesbaden: VS.

Maaz, K., Baeriswyl, F. & Trautwein, U. (2011). *Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Eine Studie im Auftrag der Vodafone Stiftung Deutschland.* Verfügbar unter http://www.vodafone-stiftung.de/publikationmodul/detail/33.html [29.06.2012].

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79* (3), 280–295.

Mullis, I. V. S. & Stemler, S. E. (2002). Analyzing Gender Differences for High Achieving Students on TIMSS. In D. F. Robitaille & A. E. Beaten (Eds.), *Secondary Analysis of the TIMSS Data* (pp. 277–290). Dordrecht: Kluwer Academic Publishers.

Neumann, M., Trautwein, U. & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. *Studies in Educational Evaluation, 37* (4), 206–217.

Neumann, M., Nagy, G., Trautwein, U. & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen. Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. *Zeitschrift für Erziehungswissenschaft, 12* (4), 691–714.

Radtke, F.-O. (2004). Die Illusion der meritokratischen Schule. Lokale Konstellationen der Produktion von Ungleichheit im Erziehungssystem. *IMISBeiträge, 23,* 143–178.

Raudenbush, S. W., Bryk, A. S. & Congdon, R. (2004). *HLM 6 for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International, Inc.

Schuler, H. (2006). Noten und Studien- und Berufserfolg. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3. überarbeitete und erweiterte Aufl.) (S. 535–541). Weinheim: Beltz.

Sperlich, A. (2009). Managementaufgabe Studierendenauswahl – Private Hochschulen als Pioniere. In M. Bülow-Schramm (Hrsg.), *Hochschulzugang und Übergänge in der Hochschule: Selektionsprozesse und Ungleichheiten.* 3. Jahrestagung der Gesellschaft für Hochschulforschung in Hamburg, 2008 (S. 71–80). Frankfurt am Main: Peter Lang.

Stanat, P. & Edele, A. (2011). Migration und soziale Ungleichheit. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung – Gegenstandsbereich* (S. 181–192). Wiesbaden: VS.

Tent, L. (2006). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3. überarbeitete und erweiterte Aufl.) (S. 873–880). Weinheim: Beltz.

Trautwein, U. & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind. Referenzgruppeneffekte bei Übertrittsentscheidungen. *Zeitschrift für Pädagogische Psychologie, 21* (2), 119–133.

Trautwein, U., Köller, O., Lehmann, R. & Lüdtke, O. (Hrsg.). (2007). *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten*. Münster: Waxmann.

Publikation 2: Vergleichbarkeit der Halbjahresnoten in Mathematik

Maué, E. (2016). Achievement—and what else? The standardisation of semester grades due to the implementation of state-wide exit examinations. Studies in Educational Evaluation, 51, 42-54.

ABSTRACT

The paper explores the shift from course-based to state-wide exit examinations at the end of upper secondary education in Germany between 2007 and 2011, and whether this resulted in an increased standardisation of teacher-assigned course-based semester grades. The sample consisted of 253 (2007) and 338 students (2011) in math courses at advanced level (schools: N = 19). Analyses of subgroups of students based on gender, ethnicity, and family background revealed a significant difference in grades. Perhaps the enhanced correlations between the achievement test and the course-based semester grades are an effect of standardisation due to state-wide exit examinations. In contrast, when achievement was controlled for, the implementation of state-wide exit examinations did not increase the standardisation of course-based semester grades in the given time in the intended manner. The course-based semester grades continue to differ depending on students' background. Several possible explanations for this result are discussed.

1. INTRODUCTION

In recent years, the education systems of many countries have begun to allow schools more autonomy to better meet their students' needs. There has further been an increase in the use of standards-based accountability systems (Woessmann, Luedemann, Schuetz, & West, 2009), although the conditions of standardised testing vary between different countries (Klein & van Ackeren, 2011). State-wide standardised tests are believed to improve educational purpose, processes, and structures to raise achievement, comparability, and equity (Bishop, 1995; Corbett & Wilson, 1991; Reardon, Atteberry, Arshan, & Kurlaender, 2009).

In Germany, a number of reforms have enlarged the level of standardisation. These include the shift from course-based to state-wide exit examinations as well as the implementation and examination of 'educational standards'. The aim of these reforms is to raise the comparability of achievement, grades, and certificates, in light of differences between classes, schools, and federal states (Maag Merki, 2014; van Ackeren, Block, Klein, & Kühn, 2012). Such comparability is particularly important with regard to the upper secondary school leaving certificate Abitur, as it constitutes the requirement to begin one's academic studies. However, it has been found that several factors such as students' background or teachers' beliefs, stereotypes, and expectations continue to influence grades. To date, neither high comparability of achievement and examination grades nor equality of opportunity have been attained (e.g. Holmeier, 2013; Maué, 2013). The question remains, therefore, how the aforementioned tests can affect grading practices to improve equity and fairness, particularly in light of the low-stakes they offer teachers and schools.

Research desideratum are thus the possible effects of the implementation of state-wide exit examinations on teacher-assigned course-based semester grades. The semester grades are of interest firstly because they constitute part of the students' grade point average alongside the grades they attain in the exit examinations. Secondly, the origin of these grades is significant: they emerge from an interactive process of teaching and learning that occurs daily between the teacher, the course as a whole, and the individual students. Following the theory of 'washback' effects (Bishop, 1995; Cheng, Watanabe, & Curtis, 2004) as well as the 'standards-based accountability theory of action' (Hamilton, Stecher, Russell, Marsh, & Miles, 2008), this paper examines whether the shift from course-based to state-wide exit examinations increases the comparability of grades in advanced-level math courses in the last four semesters before graduation. It seeks to determine the influence of individual math achievement as well as students' gender, ethnicity, and family background on these grades in the case of the German state of Bremen, where low-stakes, state-wide, upper secondary education exit examinations were implemented in 2008 in advanced-level mathematics. The database is part of a longitudinal study surveying student cohorts from 2007 to 2011. The long-term effects of this change in the examination system could be analysed by comparing grades under the conditions of course-based (2007) and state-wide exit examinations (2011).

The examination system at the end of academic-track secondary education in Germany, the theoretical and empirical background of the research as well as the research questions, hypotheses, and methods will be considered before the standardisation effects of state-wide exit examinations on course-based semester grades are analysed and discussed. Limitations, remarks for further research, and a conclusion finish this paper.

2. EXAMINATION SYSTEM AT THE END OF ACADEMIC-TRACK UPPER SECONDARY EDUCATION IN GERMANY

In academic-track upper secondary schools, the last two years before graduation are divided into four semesters with courses at basic or advanced levels. The students can choose these courses on the basis of certain regulations. Schools, teachers, and students know the general state-wide exit examination topics which will be examined in each course more than two years before the state-wide exit examinations. The general examination topics correspond to the curriculum and the examination standards of each subject and constitute a framework for the content and focus of instruction of each course. However, individual teachers are responsible for their concrete implementation (including the use of teaching materials). The teachers develop, conduct, and grade oral and written tests, homework, presentations of projects, practical skills, and so forth, in order to assess the students' achievement. The number, lengths, and requirements of the written tests is prescribed (Verordnung über die Gymnasiale Oberstufe (GyO-VO) vom 1. August 2005, §12). The single grades count towards the computation of a semester grade for each course and semester. These 'assess rather general achievement across different specific topics within one subject' (Friedrich, Flunger, Nagengast, Jonkmann, & Trautwein, 2015, p. 2). The final grade point average is comprised of the course grades in the last four semesters before graduation and the grades achieved in the exit examinations (written and oral). All grades are determined by the teacher of the course; only the exit examinations are graded by two teachers. Under course-based exit examinations, the teachers themselves develop and grade the exit examinations, whereas under state-wide exit examinations an external commission is responsible for the tests and the grading criteria. Nevertheless, the grading itself lies in the teachers' hand. According to Bishop (1995), this is not a problem as long as there are 'good rubrics for grading' (p. 680). The grades in the state-wide exit examinations have

consequences for students with regard to their futures, as they form part of the grade point average (high-stakes). However, at the macro-level of the entire school system, the results of the state-wide exit examinations have no further consequences for teachers and schools (low-stakes).

This paper focuses on the German state of Bremen. Bremen switched from course-based to state-wide exit examinations for all basic-level courses in 2007 and in 2008 for advanced-level courses in the subjects math, science, German, and foreign languages. Other subjects at advanced level still have course-based exit examinations.

3. THEORETICAL BACKGROUND

There are several reasons why the implementation of state-wide exit examinations might affect these semester grades, taking into considerations the complexity both of the multilevel structure of the educational system and of the diverse stakeholders involved, their own tasks and interests (cf. Fend, 2008).

3.1 Professional responsibility

Even if 'the accountability is low-stakes for both teacher and schools […] many teachers feel professionally responsible for their schools' results. Hence the stakes are perceived to be higher by these teachers' (Klinger & Rogers, 2011; p.140). This observation may also be valid for the German system, where test-based accountability is limited. Although low-stakes, state-wide exit examinations increase the transparency of teachers' performance by revealing whether their students fulfil external requirements. The pressure on teachers is thus greater (Bishop,1995; Woessmann et al., 2009) and may encourage them to reflect on and change their teaching practices, including their grading.

State-wide exit examinations further have the potential to shape educational perceptions or understandings and to spread innovations (Haertel, 2013; Kühn, 2011; also Cheng & Curtis, 2004). As one its aims is to increase the comparability of students' grade point average (making all grades more standardised i.e. based on students' achievements), the implementation of state-wide exit examinations might be a reason to discuss wider educational aims, for example a higher level of equity, and the steps which could be undertaken to achieve these.

3.2 Washback effects

The implementation of new examinations is accompanied by (un)intended 'washback' (or 'backwash') effects on the educational system, stake-holders, curriculum, teaching, and learning (Bishop, 1995; Cheng et al., 2004; Haertel, 2013). As noted by Cheng and Curtis (2004), washback 'refers to the influence of testing on teaching and learning' (p. 4). Prodromou (1995) defines washback more narrowly: as 'the direct or indirect effect of examinations on teaching methods' (p. 13). He distinguishes between overt (explicit) and covert (implicit) washback as well as between positive and negative (mostly overt) consequences. Watanabe (2004) points out the complexity of washback processes and differentiates between several

- dimensions: specificity (general—specific), intensity (strong—weak), length (short-term—long-term), intentionality (intended—unintended), and value (positive—negative)
- aspects of learning and teaching that may be influenced by the examination: washback to the learner (learning) or the programme (teaching)
- factors mediating the process of washback: test, personal, micro-context, and macro-context factors (p. 20ff.).

Following both Cheng and Curtis' (2004) wider definition as well as Watanabe (2004), an increase in the comparability of teacher-assigned semester grades would in this case constitute a specific, weak, intended, and positive washback effect to the programme of the implementation of low-stakes statewide exit examinations.

Cheng and Curtis (2004) suggest that 'tests or examinations can and should drive teaching, and hence learning [. . .] In order to achieve this goal, a "match" or an overlap between the content and format of the test or the examination and the content and format of the curriculum (or 'curriculum surrogate' such as the textbook) is encouraged' (p. 4).

In Germany, the content of and solutions to former state-wide exit examinations are published and used both by teachers in classroom and by students at home in preparation for the exit examinations. If the teachers integrate these items into their instruction and in consequence change their teaching

materials (Amengual Pizarro, 2010; also Cheng, 2004), it would follow that they not only use the items but also the external state-wide grading criteria for their own tests (van Ackeren et al., 2012). The assignment of course-based semester grades would thus depend more on the criterion-related reference norm than on the social or the individual ones, and less on teachers' individual characteristics (such as their expectations or stereotypes).

Furthermore, the turn toward state-wide exit examinations is intended by the regulations of the state on the last two years of upper secondary education. For this reason, teachers' tests are to be adapted step-by-step to the requirements of the written exit examinations. One of these tests must have the duration of the exit examination (Verordnung über die Gymnasiale Oberstufe (GyO-VO) vom 1. August 2005, §12).

3.3 Standards-based accountability theory of action

According to the 'standards-based accountability theory of action' (Hamilton et al., 2008), state standards and assessments, in combination with incentives, information, and assistance, influence teachers' actions directly and indirectly, mediated by the reactions of their school and district (cf. Fend's, 2008, concept of 'recontextualisation'). This process is further affected by the opinions and attitudes of several stakeholders which interact with contextual characteristics of the school, the educational system, and the society (Stevens & Görgöz, 2010). By means of a 'feedback loop', 'testing at the end of the year [leads; E.M.] to consequences in the following year' (Hamilton et al., 2008; p. 35). This implies that state-wide exit examinations no longer form the end of teaching and learning but the starting point (Cheng & Curtis, 2004).

In terms of this study, it would follow that the change in the testing system has an impact on students, teachers, and schools not only during preparation for the state-wide exit examinations but also afterwards (cf. Maag Merki, 2014). Due to higher transparency, students, teachers, and schools would focus more on the examinations themselves. This is in line with Cheng (2004) who stated that 'changing the examination is likely to change the kind of exam practice, but not the fact of the examination practice. [. . .] It might even re-focus teachers' attention on the exam' (p. 164). For instance, teachers could reflect

on the relationship between their instruction prior to the state-wide exit examinations and the contents of the examinations, their students' achievement, and their own grading in order to draw conclusions for the next course (Goldberg & Roswell, 2000; Klinger & Rogers, 2011). It would be a washback effect if the teachers were to adopt the external state-wide grading criteria of one year's state-wide exit examinations for their own course-based examinations in the following year(s) and use them to derive semester grades. Such grading on the criterion-related reference norm rather than the social one could lead to higher standardisation and comparability of semester grades and therefore of the grade point average. The German federal state of Bremen 'implemented a monitoring process by which after the exit exam, some of the exam grades are reanalyzed by responsible professionals at the state level for monitoring purposes' (Maag Merki & Holmeier, 2015; p. 60). At that stage, specialists provide feedback to teachers on the accuracy of their grading (ibd., p. 61). A high accuracy can only be reached if the grades reflect the students' achievement and not other factors, as for example students' background or behaviour. As well as the grades, the exit examinations in its entirety are analysed by an examination commission in cooperation with some teachers. The results are discussed in the schools and conclusions are drawn for the next cycle of exit examinations.

This monitoring process can be seen as one part of a feedback loop which might facilitate washback effects and changes in educational processes. Furthermore, the teachers' individual and collective experiences with the state-wide exit examinations themselves, grading criteria, grading process, and external feedback as well as with their participation in the examination commissions are important factors (Black, Harrison, Hodgen, Marshall, & Serret, 2011; Goldberg & Roswell, 2000; Hofer, 2015).

'The adaptation process between stronger standardization of the grading is additionally supported by the fact that the members of the examination commissions in the German states are teachers in active service, and their experiences flow back into school practice in their own schools' (Maag Merki & Holmeier, 2015; p. 61).

3.4 Teachers' beliefs, perceptions, stereotypes, and expectations

In the theories mentioned above, the beliefs, assumptions, and knowledge-levels of different stakeholdes are important factors in examining changes in educational processes. Analysing the students'

grades is related to the complex interplay of teachers' and students' interactions in class. This implies that theories about teachers' beliefs about education, instruction, and students have to be taken into account. According to a theoretical and empirical review from Fives and Buehl (2012), teachers hold beliefs about their self, context or environment, content or knowledge, specific teaching practices and teaching in general as well as about students (p. 472). Teachers' beliefs are organized into one or more system(s) of implicit and explicit, stable and dynamic beliefs which are interwoven with teachers' knowledge and context demands (ibd., p. 473ff.). These beliefs fulfil several functions: they serve as a filter for information and experience, frame situations and problems, and guide intentions and actions (ibd., p. 478ff.). Teachers' beliefs may 'affect how they interpret pedagogical reforms (filter) and what they perceive as the task at hand (frame)' (ibd., p. 479). In the case of this paper, teachers must recognize a higher standardisation of grades as an important aim of the implementation of state-wide exit examinations and draw conclusions for their actions (grading), based on beliefs about teaching and learning which may change over time and with experience.

Teachers' stereotypes and expectations are further related to their judgement of students' achievement. Jussim, Eccles, and Madon (1996, p. 296f.) have developed a model which assumes that students' achievement is influenced by the students' background as well as the teachers' perceptions (= self-fulfilling prophecies), which in turn are affected by the students' background and other possible moderating factors (e.g. students' previous achievement, gender, ethnical and socioeconomic background). Teachers' perceptions and such self-fulfilling prophecies may be biased by teachers' stereotypes, attitudes, and expectations (Jussim et al., 1996; van Ewijk, 2011) which may result in different behavioural practices and treatment (Rosenthal & Jacobson, 1992; Tenenbaum & Ruck, 2007). The impact of teachers' expectations on students' intelligence and achievement is also known as 'Pygmalion effect': 'one person's expectation for another's behavior could come to serve as a self-fulfilling prophecy' (Rosenthal & Jacobson, 1992, p. 174). Students' awareness of differential teachers' expectations, beliefs, and behaviour can lead to poorer students' achievement (e.g. caused by stereotype-threat), which in turn panders to teachers' expectations or stereotypes ('carryover' Weinstein, 2002; also Forghani-Arani, Geppert, & Katschnig, 2015).

In this case, the teachers' beliefs, (in)accurate stereotypes, perceptions, and expectations may affect their students' grades.

In sum, the theory of washback effects of educational reforms (e.g. Bishop, 1995; Cheng et al., 2004) and the 'standards-based accountability theory of action' (Hamilton et al., 2008) form the broader theoretical framework for an explanation of effects of educational reforms on teaching and learning, in this case, the possible impact of the implementation of state-wide exit examinations on teacher-assigned semester grades. If the state-wide grading criteria, as one aspect of the reform, fit to or change the teachers' beliefs and guide their actions (grading), their expectations and judgements should be based on students' achievement rather than on inaccurate stereotypes related to students' background (Fives & Buehl, 2012; Jussim et al., 1996; van Ewijk, 2011). This, in turn, should result in a higher standardisation of grades.

4. PREVIOUS RESEARCH

4.1 Impact on grades

The reliability and validity of grades has often been criticized. As early as 1974, Gronlund stipulated 'letter grades and the check lists of objectives should reflect achievement and achievement only. They should not be contaminated by the teacher's judgement of student effort or student behaviour' (p. 13). Nonetheless, research has shown the opposite: grades were influenced by a combination of teacher-, class-, and student-related factors.

The impact of *teachers'* stereotypes, attitudes, perceptions, and expectations on students' achievement is well documented. For instance, Friedrich et al. (2015) demonstrated that teachers' expectations affected the students' grades and test scores at the individual but not at the classroom level (controlled for several students' characteristics; also Trouilloud, Sarrazin, Martinek, & Guillet, 2002).

Jussim et al. (1996) provided evidence for the accuracy of teachers' perceptions, stereotypes, and judgement of their students. To a great extent, teachers' perceptions of similarities and differences in performance (grades) and talent (standardised test scores) depending on the students' gender, social class, and ethnicity was accurate (with one exception). The results of teachers' perception of the students' effort were mixed and related to the measurement of effort (p. 329ff.). Furthermore, when controlled for

actual achievement and motivation, students' socioeconomic and ethnical background had no influence on teachers' judgements. With regard to students' gender, the results ranged from no stereotype in terms of talent to accurate stereotype in terms of performance, and inaccurate stereotype in terms of effort (ibd., p. 345ff.). Moreover, teachers' expectations had a greater effect on girls, students with a lower socioeconomic background, and African-American students. It also turned out that 'students with multiple vulnerabilities are more susceptible to self-fulfilling prophecies than […] students with only one vulnerability' (ibd., p. 359; Weinstein, 2002).

Teachers' expectations differed depending on students' individual characteristics: girls and students with lower grades perceived lower teacher ability expectations (Lazarides & Watt, 2015; Weinstein, 2002). Following Tenenbaum and Ruck's meta-analysis (2007), teachers had higher expectations, more positive and fewer negative referrals as well as more positive and neutral speech towards European American students than towards minority students (similar: Rubie-Davis, Hattie, & Hamilton, 2006). This is in line with van Ewijk (2011) who found that ethnic majority teachers did not exhibit a direct grading bias concerning ethnic minority and majority students. However, different behaviour towards ethnic minority students based on lower expectations and negative attitudes did have an indirect effect (contrary results: Forghani-Arani et al., 2015). Stevens and Görgöz (2010) observed that teachers' view of underachieving ethnic minority students was related to the examination system: if it lay in the teachers' hands, as is the case in Belgium, teachers 'were more likely to blame Turkish students for not reaching appropriate standards imposed by their teachers' (p. 1361). In contrast, teachers in England with standardised external examinations 'tend to view these students as victims of a situation that has left them largely unequipped to meet externally imposed standards' (ibd.).

Beyond that, several *sources of error* due to teachers' attitudes during the grading process, such as positive/negative errors of leniency, errors of contrast, errors of ranking, halo-effects, and effects of severity or leniency might distort the grades (e.g. Hochweber, 2010). In addition, the *reference norm of grading* (individual development, social comparison, or criterion-related) is important and related to teachers' beliefs (Rakoczy, Klieme, Bürgermeister, & Harks, 2008). Tierney, Simon, and Charland (2011) found that teachers agreed to criterion-referenced grading on the one hand, and on the other varied the frame of reference in dependence of the students. Especially, when teachers based their

judgements on social comparisons, the average achievement of the *reference group*, in most cases the *class*, influenced the grading, also known as 'Big-Fish-Little-Pond-Effect' (e.g. Dardanoni, Modica, & Pennisi, 2011; Marsh & O'Mara, 2010; Møen & Tjelta, 2010; Neumann, Nagy, Trautwein, & Lüdtke, 2009; Neumann, Trautwein, & Nagy, 2011).

In line with the evidence for the meaning of teachers' characteristics, research has shown that grades depend not only on students' achievement but also on *non-cognitive factors* such as their motivation, self-concept, involvement, social and behavioural aspects or, as Bowers stated, 'the social processes of schooling' (2011, p. 141; Hochweber, 2010; Klapp Lekholm & Cliffordson, 2009; Trouilloud et al., 2002).

Besides, but not necessarily separate from these factors, *students' background* plays an important role. With regard to students' *gender*, research findings mostly revealed an advantage for females despite same or lower achievement (e.g. Cappellari, Lucifora, & Pozzoli, 2012; Klapp Lekholm & Cliffordson, 2009; Maaz, Baeriswyl, & Trautwein, 2011; contrary results: Hofer, 2015). Another factor which might have an impact on grades is students' *ethnicity*. In the German context, research findings concerning grading and treatment in schools extended from (institutional) discrimination to preferential treatment (e.g. Gomolla & Radtke, 2009; Klieme, 2003; Klieme et al., 2010; Maaz et al., 2011; Schräpler & Weishaupt, 2013). However, in many cases, ethnicity is not of prime importance, but rather differences in parental education and students' *socioeconomic family background* (e.g. Bornkessel & Kuhnen, 2011; Boykin, 2010; Maaz et al., 2011; Resh, 2010; Thorsen, 2012).

Altogether, these studies provide evidence that grades are influenced by a multitude of factors and noncognitive student characteristics, indicating the continual relevance of Gronlund's (1974) demand that grades should only assess students' achievement.

4.2 Grades and state-wide exit examinations

State-wide exit examinations are seen as an effective instrument to improve not only the comparability of grades but also other aspects as well (Maag Merki, 2014).

'A system of state or national curriculum-based exams would remedy the problem of noncomparable standards and grading relative to others in the school. It would generate a set of reliable signals of absolute levels of achievement in high school' (Bishop, 1995; p. 684).

According to van Ackeren et al. (2012), the topics and grading criteria of state-wide exit examinations had an influence on the objectivity in grading particularly in math.

With regard to the relation between grades in high school and achievement tests, research indicated standardisation effects of standards-based grading and state-wide exit examinations (Haptonstall, 2010). Paeplow (2008, 2011) found that the correlations between standards-based grades and end-of-grade standardised achievement tests in elementary schools were stronger than the correlations between non-standardised grades and end-of-grade standardised achievement tests in middle schools. Over a three year time-span, the relation between standards-based grades and the achievement test narrowed.

The findings of this study went in the same direction, showing that, concerning the four semester grades before graduation, the shift to state-wide exit examinations increased the comparability of the semester grades (Maag Merki & Holmeier, 2015). However, these analyses only focussed on the relationship between the semester grades and the math achievement test and did not include students' background. Other studies with focus on math courses at advanced level in Germany revealed that, in comparison to the state-wide exit examination grades, the semester grades corresponded less to students' achievement and more to their background (Maaz et al., 2011; Neumann et al., 2009, 2011).

Research findings concerning the effects of state-wide exit examinations on semester grades are limited. However, studies about the grades in the examinations and the final grade point average have been conducted. Schräpler and Weishaupt (2013) compared students' grade point average at the end of upper secondary education under course-based and state-wide exit examinations in the German federal state of North Rhine-Westphalia. They offered important insights about differential effects on grades depending on students' gender and ethnicity, although they did not include data about students' achievement or cognitive competences to analyse possible discriminations in grades. For math courses at advanced

level, previous findings of this study investigated that, as a short-term effect (2007–2008), state-wide exit examinations minimised the influence of students' ethnicity but intensified the dependence of the grades in the exit examinations on the students' family background. In a long-term perspective (2007–2011), however, state-wide exit examinations did not improve the comparability of grades in the exit examinations in math courses at advanced level (Maué, 2013; cf. Holmeier, 2013).

Additionally, other analyses with cross-sectional data from math and science teachers in the German federal state of Bremen suggested an increase in the use of criterion-related reference norm of grading in courses with state-wide exit examinations from 2007 to 2008 and 2009 (Holmeier, 2013).

All in all, it can be stated that research findings concerning the comparability of grades in state-wide exit examinations differ depending on the analysed subject, course level, and region. Furthermore, the effects of non-achievement factors on semester grades in the last four semesters before graduation have not been sufficiently analysed to date.

5. PURPOSES

Although several questions concerning the comparability of grades have been answered, there is a lack of longitudinal studies analysing possible washback effects of state-wide exit examinations on the course-based semester grades in the four semesters before graduation. Since they are part of the students' grade point average, these grades constitute the focus of interest. Based on the aforementioned theory of washback effects (Bishop, 1995; Cheng et al., 2004) and the 'standards-based accountability theory of action' (Hamilton et al., 2008), this paper investigates whether state-wide exit examinations affect grading practices. Therefore, the effects of students' background (gender, ethnicity, and family background) on the teacher-assigned semester grades are compared under the conditions of course-based (2007) versus state-wide exit examinations (2011). As state-wide exit examinations were implemented in 2008, it is possible to analyse long-term effects.

First, the following research question shall be answered: *Do the semester grades differ between various subgroups (gender, ethnicity, and family background) in the German state of Bremen?*

In line with research citing differences in grades favouring certain groups of students, better grades are hypothesized for female students (e.g. Klapp Lekholm & Cliffordson, 2009; Maaz et al., 2011; Resh, 2010), students who were born in Germany (e.g. Maaz et al., 2011; Schräpler & Weishaupt, 2013), and students who have more books at home (e.g. Bornkessel & Kuhnen, 2011; OECD, 2012).

Second, the following research questions are in the focus of interest: *Do state-wide exit examinations in the German state of Bremen increase the comparability of grades in math courses at advanced level in the four semesters before graduation? Specifically, will students get similar semester grades, independent from their gender, ethnicity, and family background when controlled for individual achievement?*

Based on the assumption that standardised testing might be an instrument to improve educational purpose, processes, and structures to raise comparability and equity (Bishop, 1995; Corbett & Wilson, 1991; Maag Merki, 2014; Reardon et al., 2009) and the higher standardisation of state-wide exit examinations in comparison to course-based ones (e.g. Haptonstall, 2010; Maag Merki & Holmeier, 2015; Paeplow, 2011), it is hypothesized that semester grades, under the conditions of state-wide exit examinations, depend more on students' individual achievement. Concomitant with this, the effects of gender, ethnicity, and family background on semester grades should diminish from 2007 (course-based) to 2011 (state-wide exit examinations).

6. METHOD

6.1 Data

Data comes from a longitudinal study which analysed the shift from a course-based to a state-wide organized exit examination system at the end of academic-track upper secondary education during 2007 to 2009 and in 2011. The study takes differing perspectives (students, teachers, and school administrators) as well as multi-dimensional criteria (school quality, diagnostic processes, teaching quality, self-regulated learning, and students' achievement) into account. In the years 2007, 2008, 2009, and 2011, achievement tests in math and English, standardised surveys for teachers and students, evaluations of examinations and grading data, as well as a qualitative case analysis were conducted in the German

federal states of Bremen and Hesse (Maag Merki, 2012). In each school in the sample, all students of one math and one English course at basic and at advanced level were chosen for the participation in the study. At the individual level, the sample consists of cross-sectional data, whereas at the school level it is a longitudinal study.

For this paper, only math courses at advanced level in the German federal state of Bremen are of interest, due to the fact that data from course-based (2007) as well as from state-wide exit examinations (2011) is available. In 2007, all semester grades were assigned under the conditions of course-based exit examinations, whereas in 2011 they all were created under the conditions of state-wide exit examinations. In 2008 and 2009 the grades derived under a mixture of both conditions leaving 2011 as the only year with this data.

6.2 Sample

The *sample* in the German state of Bremen includes all upper secondary schools (N = 19). The response rate for students increased from 51% in 2007 (n = 751) to 74% in 2011 (n = 1157).

For the following analyses, the sample consists of 253 (2007) and 338 students (2011) from each one math course at advanced level per school. All students in the analyses took math as written exit examination subject at advanced level. Due to the limitation that only courses with at least five students were taken into consideration for the multilevel analyses, the sample was reduced to 12 schools per year (n = 24) with 180 (2007) and 215 students (2011). Although the sample size is small, multilevel analyses can be computed because, according to Maas and Hox (2005), the 'estimates of the regression coefficients are unbiased, even if the sample is as small as 10 groups of five units' (p. 91). However, the estimates of the standard errors of the second level might by biased. Concerning the distribution of students' background (gender, ethnicity, and family background), t-Tests reveal no significant differences between the students of the schools included in and excluded from the analyses in 2007 and in 2011 (gender: 2007: $t(200) = -1.469$, $p = 0.143$; 2011: $t(280) = -0.159$, $p = 0.874$; ethnicity: 2007: $t(146) = -0.688$, $p = 0.492$; 2011: $t(262) = 0.007$, $p = 0.995$; books at home: 2007: $t(148) = 0.979$, $p = 0.329$; 2011: $t(258) = 1.312$, $p = 0.191$). The selected students do not differ between 2007 and 2011 in the distribution of gender ($t(392) = -0.764$, $p = 0.445$) and of ethnicity ($t(345) = 0.249$, $p = 0.803$) but the

students in 2011 have significantly more books at home than in 2007 ($t(344) = 3.628$, $p < 0.001$; $d = -0.39$). However, all in all the sample is comparable from 2007 to 2011 and is representative for the German state of Bremen (Table 1).

Table 1: Distribution of missing values and individual characteristics of the sample

	2007		2011	
Variables	N (%)	% missing	N (%)	% missing
N max	253		338	
Semester Grades				
12/1	249	1.6	334	1.2
12/2	249	1.6	333	1.5
13/1	249	1.6	334	1.2
13/2	249	1.6	334	1.2
Math test	204	19.4	295	12.7
Gender	239	5.5	338	
Female	89 (37.2)		102 (30.2)	
Male	150 (62.8)		236 (69.8)	
Country of origin	163	35.6	319	5.6
Born in Germany	141 (86.5)		276 (86.5)	
Not born in Germany	22 (13.5)		43 (13.5)	
Family background	165	34.8	315	6.8
Less books at home	75 (45.5)		191 (60.6)	
More books at home	90 (54.5)		124 (39.4)	

6.3 Instrument and variables

As instrument 15 items of the achievement test 'TIMSS/III Advanced Mathematics' (Klieme, 2000) are used. For information about reliability and validity of the achievement test see Klieme (2000) as well as Maag Merki and Holmeier (2015). Kahnert (2014) showed for North Rhine-Westphalia that this test and the test items of the state-wide exit examination in math measure the same general mathematical competence. It can thus be assumed that students take the test seriously and perform at their actual level, although the test is low-stakes for them. As an additional exercise before the real examination, it contains the possibility of seeing what they already do or do not know. Each school received its test results in comparison to the years before and to all schools, as well as for all students who took the test and for the students who had math as an examination subject (no feedback on individual basis). It has to remain unanswered in what way the schools use this information to change teaching and learning.

Furthermore, the semester grades, a combination of oral and written tests, in the last four semesters before graduation (semesters 12/1, 12/2, 13/1, and 13/2; see description of the German system) are a focus of interest. The grades, as well as the achievement test, have a range of 0 to 15 points. The grading scale can be transformed to the six grades usually used in Germany: 0 points = fail (grade 6), 1–3 point(s) = below average (grade 5), 4–6 points = sufficient (grade 4), 7–9 points = satisfactory (grade 3), 10–12 points = good (grade 2), 13–15 points = excellent (grade 1). The students' background is specified by the variables gender (0 = female, 1 = male) and country of origin (0 = Germany, 1 = foreign country). Although this is a narrow indicator for the students' ethnicity (see limitations), it may give hints about differential treatments of teachers depending on students' migration background. It distinguished between students who were born in Germany and visited all educational institutions there and those who immigrated and potentially also visited educational institutions in their country of birth prior to the upper secondary school in Germany. The students' family background is indicated by the number of books at home (0 = 0–10 to 5 = more than 500). For a better clarity, for the basic analyses a dichotomous version of the variable number of books at home (0 = below mean, 1 = at or above mean) is used. For the multilevel analyses, the six-tier variable is employed. It is proved that the number of books at home is a good indicator for the social status of students' families.

6.4 Analyses

To cover the complexity of the data and the research questions, diverse *analyses* are used. Besides descriptive analyses, differences between subgroups (female versus male students, Germany versus other country of origin, and less versus more books than the average at home) as well as changes in the semester grades between the years 2007 and 2011 are determined by t-Tests as well as effect sizes (Cohen, 1988). Changes in the correlations between the math achievement test and the four semester grades over time are proved by Fishers Z-transformation. To be faithful to the multilevel structure of educational systems (Fend, 2008), multi-level analyses are computed with each semester grade as a dependent variable. The result in the math achievement test, gender, country of origin, and number of books serve as independent variables. All variables are not centred, except for the achievement test, which is centred around its grand mean at both levels. The final model is each time an intercepts and slopes-as-Outcomes Model (in analogy to Holmeier, 2013; Maué, 2013), computed with HLM 6.06 (Raudenbush, Bryk, & Congdon, 2004). Listwise deletion of missing data at individual level was conducted when running the analyses. At individual level, the impact of these variables on students' grades is examined. At course level, possible effects of the level of achievement of the course as well as changes in the grades over the years are modelled. By means of interaction effects between the year 2011 and the result in the math achievement test, gender, country of origin, and number of books, it is proved whether the implementation of state-wide exit examinations changed the effects of these variables on the semester grades. Data from the year 2007 (course-based exit examinations) serves as the reference year so that the comparison between 2007 and 2011 depicts the long-term effects of state-wide exit examinations on course-based semester grades.

$$grade = \gamma 00 + \gamma 01 * Achievement\ Test\_Level2 + \gamma 02 * Year2011 + \gamma 10 * Gender + \gamma 11 * (Year2011 * Gender) + \gamma 20 * Country\ of\ origin + \gamma 21 * (Year2011 * Country\ of\ origin) + \gamma 30 * Books\ at\ home + \gamma 31 * (Year2011 * Books\ at\ home) + \gamma 40 * Achievement\ Test\_Level1 + \gamma 41 * (Year2011 * Achievement\ Test\_Level1) + u0 + u1 * Gender + u2 * Country\ of\ origin + u3 * Books\ at\ home + u4 * Achievement\ Test\_Level1 + r$$

7. RESULTS

7.1 Semester grades differentiated by students' background

Table 2 presents an overview of the descriptive statistics for the semester grades in the years 2007 (course-based exit examinations) and 2011 (state-wide exit examinations) for all students, further distinguished between subgroups.

Table 2: Descriptive statistics of semester grades

Grade	12/1				12/2				13/1				13/2			
	n	*M*	*SD*	*d*	*n*	*M*	*SD*	*d*	*n*	*M*	*SD*	*d*	*n*	*M*	*SD*	*d*
2007																
Female	89	9.70	2.98	-0.18	89	9.60	2.95	-0.08	89	9.64	3.20	**-0.23**	89	9.99	3.29	-0.17
Male	149	9.14	3.09		149	9.34	3.03		149	8.89	3.19		149	9.43	3.26	
Germany	140	10.18	2.81	**-1.03**	140	9.82	3.09	**-0.54**	140	9.44	3.43	**-0.43**	140	10.22	3.36	**-0.49**
Other country	22	7.23	3.19		22	8.18	2.54		22	8.00	2.51		22	8.59	2.97	
Less books	74	9.39	2.86	0.25	74	9.03	2.98	**0.34**	74	8.91	3.20	0.18	74	9.41	3.48	**0.34**
More books	90	10.16	3.13		90	10.06	3.08		90	9.51	3.42		90	10.54	3.15	
total	249	9.59	3.01		249	9.67	3.00		249	9.33	3.22		249	9.82	3.19	
2011																
Female	98	9.37	2.84	-0.04	98	9.15	3.27	0.04	98	8.99	3.36	0.02	98	9.54	3.33	-0.10
Male	236	9.24	2.99		235	9.26	3.08		236	9.06	3.16		236	9.19	3.49	
Germany	273	9.41	2.95	-0.25	272	9.41	3.12	**-0.28**	273	9.05	3.28	0.07	273	9.39	3.42	-0.04
Other country	42	8.67	2.87		42	8.52	3.12		42	9.26	3.03		42	9.26	3.55	
Less books	189	8.85	2.80	**0.43**	189	8.71	3.06	**0.49**	189	8.45	3.14	**0.50**	189	8.80	3.47	**0.42**
More books	122	10.07	3.02		121	10.20	3.02		122	10.03	3.19		122	10.24	3.22	
total	334	9.26	2.97		333	9.26	3.04		334	9.08	3.13		334	9.29	3.36	

Note. Grades: 0-15 points; 0 points = fail, 1-3 point(s) = below average, 4-6 points = sufficient, 7-9 points = satisfactory, 10-12 points = good, 13-15 points = excellent
d = positive: higher grades for male students, students not born in Germany, and students with more books;
d = negative: higher grades for female students and German students.
Bold: $p < .10$

Irrespective of the level of achievement and the examination system, there were no gender-specific differences in the semester grades in 2007 and 2011. The only exception is the semester 13/1 in 2007 where female students received in tendency a significantly higher grade.

In 2007, the students who were not born in Germany were assigned significantly lower grades in all semesters, whereas this only was the case in the semester 12/2 in 2011. In all others semesters in 2011, there were no differences in the grades between German-born and not German-born students.

The comparison of students with less and with more books than the average at home revealed that the latter group got significantly higher grades in 2007 in the semesters 12/2 and 13/2 and in 2011 in all four semesters.

By means of t-Tests and effect sizes, possible changes of the semester grades over the years were proven (12/1: $t(581) = 0.365$, $p = 0.715$; 12/2: $t(580) = 0.955$, $p = 0.340$; 13/1: $t(581) = 0.664$, $p = 0.507$; 13/2: $t(581) = 1.298$, $p = 0.195$). The t-Tests as well as the effect sizes revealed no significant and relevant changes (12/1: $d = -0.03$; 12/2: $d = -0.08$; 13/1: $d = -0.06$; 13/2: $d = -0.11$).

7.2 Relation between semester grades and students' achievement

As a first step to prove changes due to the implementation of state-wide exit examinations, correlations were calculated (Table 3).

Table 3: Correlations of math achievement test with each semester grade

Grade semester	12/1	12/2	13/1	13/2
		2007		
Math test	.23***	.31***	.27***	.31***
		2011		
Math test	.45***	.44***	.46***	.45***

Note. 2007: n = 200; 2011: n = 291 (12/2: n = 290).
*** $p < .001$

The correlations between the achievement test and the grades in the four semesters ranged from $r = 0.23$ ($p = 0.001$) to $r = 0.31$ ($p < 0.001$) in 2007 and from $r = 0.44$ ($p < 0.001$) to $r = 0.46$ ($p < 0.001$) in 2011. When comparing the correlations over the years with Fisher's Z-transformation, all relationships narrowed significantly (12/1: $p = 0.003$; 12/2: $p = 0.051$; 13/1: $p = 0.009$; 13/2: $p = 0.038$) which might be carefully interpreted as an effect of higher standardisation under the conditions of state-wide exit examinations than with course-based exit examinations.

7.3 Comparability of semester grades

As a second step, multilevel analyses were computed with each semester grade as a dependent variable to prove the influence of different factors on these grades at individual level over time (Table 4).

Table 4: Multilevel analyses for each semester grade (fixed effects with robust standard errors)

Grade semester	12/1	12/2	13/1	13/2
Fixed Effects				
Intercept	10.11*** (0.64)	10.48*** (0.82)	10.77*** (0.99)	10.25*** (0. 83)
Level 1 (students)				
Gender	-1.11^{+} (0.54)	-0.62 (0.42)	-1.41* (0.51)	-1.09* (0.49)
Country of origin	-2.39** (0.65)	-1.40** (0.44)	-0.82^{+} (0.47)	-1.16^{+} (0.65)
Books at home	0.28^{+} (0.14)	-0.00 (0.15)	-0.05 (0.16)	0.23 (0.18)
Math test	0.50*** (0.07)	0.60*** (0.08)	0.65*** (0.10)	0.60*** (0.11)
Level 2 (course)				
Math test	-0.39** (0.11)	-0.40** (0.12)	-0.51** (0.15)	-0.60*** (0.12)
Year2011	-1.64* (0.74)	-2.33* (1.01)	-2.60* (1.14)	-1.32 (0.98)
Interaction effects				
Year2011* Gender	1.01 (0.66)	0.88 (0.64)	1.46* (0.63)	0.69 (0.66)
Year2011* Country of origin	2.22* (0.84)	0.79 (0.64)	1.17^{+} (0.61)	0.80 (0.95)
Year2011* Books at home	0.06 (0.17)	0.38* (0.18)	0.44* (0.19)	0.07 (0.21)
Year2011* Math test	0.04 (0.10)	-0.13 (0.09)	-0.08 (0.12)	0.07 (0.14)
Variance components				
u0	0.99	2.73^{+}	3.57*	1.21
u1 (Gender)	0.94	0.73^{+}	0.61	0.87
u2 (Country of origin)	0.73	0.10	0.45	1.34
u3 (Books at home)	0.04	0.01	0.02	0.01
u4 (Math test)	0.02	0.01	0.03^{+}	0.04
R	5.70	5.97	6.22	7.55
Intraclass-Correlation	*0.10*	*0.12*	*0.11*	*0.08*

Note. Standard errors are in parentheses.
$^{+}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$
Grades and math achievement test: 0-15 points; gender: 0 = female, 1 = male; country of origin: 0 = Germany, 1 = foreign country; books at home: 0 = 0-10 to 5 = more than 500; year2011: 0 ≠ 2011, 1 = 2011

The comparison of the grades shows obvious similarities and differences with regard to the impact of various factors as well as their development over time. In contrast to the results of the t- Tests mentioned afore, the multilevel analyses indicate a significant decrease of the grades in the semesters 12/1, 12/2, and 13/1 from 2007 to 2011 whereas the grade in the semester 13/2 underwent no significant changes. In 2007, achievement had an important impact: students who had better results in the achievement test gained higher semester grades. However, though they had the same results in the achievement test, male students (exception semester 12/2) and students who were not born in Germany were assigned lower grades (partly in tendency). Having more books at home was in tendency only an advantage in the semester 12/1 and did not matter in the other semesters. Furthermore, the level of the reference group played a role: despite equal results in the achievement test, students in high achieving courses were assigned worse grades than students in lower achieving courses et vice versa.

The relationship between achievement and grades did not change over time (2007–2011). The same applies to gender, such that male students still faced disadvantages. The only exception is the semester 13/1, in which the disadvantage of male students was equalized. In terms of grades, students who were not born in Germany were disadvantaged in 2007 (partly in tendency) which also was the case in 2011 for the semesters 12/2 and 13/2. In contrast, the differences in the grades in the semesters 12/1 and 13/ 1 were (almost) equalized. An unintended effect occurred for the family background as students with more books at home were assigned higher grades in the semesters 12/2 and 13/1. This effect was in tendency already seen in 2007 for the semester 12/1 and continued to last (no significant interaction effect), so that only the grade in the last semester before graduation (13/2) was not influenced by the students' family background from 2007 to 2011.

In sum, the analyses of all students with math examinations at advanced level reveal that, irrespective of the level of achievement, there existed more differences in the grades between female and male students and between German-born and not German-born students under the conditions of course-based exit examinations in 2007 than under the conditions of state-wide exit examinations in 2011. The opposite is true when comparing students with less and more books than the average at home. The correlations between the grades in the four semesters and the achievement test were closer under the conditions of state-wide exit examinations which might be carefully interpreted as an effect of higher

standardisation. The t-tests indicated that the mean scores of all the semester grades remained stable from 2007 to 2011. In contrast, multilevel analyses showed that, in a long-term perspective, the grades decreased (exception semester 13/2). While the impact of achievement and gender (exception semester 13/1) on the grades did not change over time, the contrast applies to the influence of ethnicity and family background whose development varied between the different semester grades.

8. DISCUSSION

The purpose of this paper was to investigate whether the shift from course-based to state-wide exit examinations in the German federal state of Bremen led to changes in teacher-assigned semester grades, especially in terms of their comparability. To this end, semester grades in math courses at advanced level in the last four semesters before graduation were analysed within the time period of 2007 to 2011. It was assumed that, due to the higher standardisation of state-wide exit examinations, the semester grades would depend more on students' achievement and less on factors unrelated to achievement than with course-based exit examinations.

8.1 Impact of students' background and achievement on the semester grades

Prefixed is an overview of the semester grades for different subgroups and over time. It was hypothesized that female students received better grades than male students. Without controlling for achievement, this occurred neither in 2007 under the conditions of course-based exit examinations (exception semester grade 13/1) nor in 2011 under state-wide exit examinations. When achievement was controlled for, the data reveal an advantage for female students (exception semester grade 12/2) in 2007 which in 2011 turned back significantly only for the semester grade 13/1. The others did not, so that the higher grades for female students (semester grade 12/1 and 13/2) as well as no gender-specific differences in the semester grade 12/2 continued. All in all, the effect of students' gender varied over time and between the semesters. The higher grades found despite same achievement in some semesters for female students match those observed in other studies (e.g. Cappellari et al., 2012; Klapp Lekholm & Cliffordson, 2009; Maaz et al., 2011; Resh, 2010; Thorsen, 2012; Thorsen & Cliffordson, 2012).

One possible explanation for female students' advantage over male students might be the finding of Klapp Lekholm and Cliffordson (2009) that students' motivation, parental engagement, and self-perception had a connection with students' gender which in turn influences students' achievement. Another explanations might be that female students have mastered the 'social processes of schooling' (Bowers, 2011; p. 141; e.g. Resh, 2010) better than their male classmates or that teachers hold gender- specific stereotypes of students' effort (Jussim et al., 1996) which are incorporated into the grades. Entwined with the allocation of grades are emotions like (in)justice and (un)fairness in students as well as with teachers (Resh, 2009). Male students experience a higher level of injustice than female students which is accompanied by the unequal distribution of grades by gender in relation to achievement (Resh, 2010). Due to the fact that emotions are an essential part of learning and teaching (e.g. Hargreaves, 1998), the effects of assessments and grades on emotions, motivation, and effort (e.g. Brookhart, 1997; Bürgermeister, 2014) should not be underestimated.

The hypothesized disadvantage of students born in a foreign country concerning their grades existed under the conditions of course-based exit examinations in 2007 whether or not achievement was controlled for. In this area, results from 2011 under state-wide exit examinations varied. When achievement was not controlled for, there were no differences between German-born and not German-born students (exception semester 12/2). When achievement was controlled for, the disadvantage of students born in a foreign country turned back. This effect became significant and (almost) equalized the lower grades despite same achievement for the grades in the semesters 12/1 and 13/1 (in tendency) only. The others did not change significantly so that German-born students still received higher grades in the semesters 12/2 and 13/2 (in tendency).

The differences between the semester grades concerning the effect of students' ethnicity reflect the strong variation in previous research findings. On the one hand, the dependence of semester grades of students' ethnicity is in line with findings of several studies (e.g. Maaz et al., 2011; Weishaupt & Schräpler, 2013). On the other hand, semester grades not differing between German-born and not German-born students confirm the analyses of Bornkessel and Kuhnen (2011), Jussim et al. (1996) or van Ewijk (2011). Only the 'positive discrimination' (e.g. Klieme, 2003; Klieme et al., 2010) of foreign students could not be shown. Keeping in mind that students who were not born in Germany were assigned

lower grades despite the same achievement in 2007, the partial equalization of the discrimination in 2011 might be interpreted as a standardisation effect of state-wide exit examinations, at least on some semester grades. However, it must be noted that the students' ethnicity was based on their country of birth which is a narrow definition. The results might be different if other aspects, for instance the language spoken at home, would have been taken into consideration.

Students with less *books at home* than the average were assigned lower grades in 2007 (only in the semesters 12/2 and 13/ 2) and in 2011, independent of their achievement. When achievement was controlled for, students' family background had no impact on grading in 2007 (exception semester 12/1 in tendency). This is in contrast to previous research (e.g. Bornkessel & Kuhnen, 2011). However, when taking the long-term perspective into consideration, it becomes evident that the grades were influenced by the students' family background (exception semester 13/2) so that, all in all, the hypothesis of lower grades for students with less books at home is validated. Further analyses must investigate the unintended effect of the last observed year. The result that students with a higher social status received better grades is in line with findings from Thorsen (2012) for norm-referenced grades, whereas in contrast, criterion-referenced grades favoured students with lower social status (Thorsen & Cliffordson, 2012). The latter was not the case in this study.

The correlations between the grades in the four semesters and the achievement test have significantly enhanced since the implementation of state-wide exit examinations. In 2011, they were comparable to the correlation between the grade in the math exit examination and this achievement test (Kahnert, 2014; cf. Hochweber, 2010), although slightly lower than reported from Neumann et al. (2009, 2011). This narrowed relationship might be interpreted as an increase in the importance of students' achievement for grades. It seems, in line with the hypothesis, that the higher standardisation of state-wide exit examinations (p. e. external state-wide grading criteria) thus influenced the course-based grades in the four semesters before graduation. In contrast to this, the results of the multilevel analyses indicated no increased impact of students' achievement on the grades under the conditions of state-wide exit examinations (no significant interaction effect 2011) when achievement was controlled for. Findings of the same study with data for the years 2007, 2008, and 2009 (Holmeier, 2013) as well as 2007, 2008, and 2011 respectively (Maué, 2013) also showed no increase in the influence of students'

achievement on grades in the math exit examinations over time. Based on this, the hypothesis concerning a stronger orientation of grades by students' achievement under state-wide exit examinations over time must be refuted.

All in all, it has to be stated that, in general, the implementation of state-wide exit examinations did not lead to an increased standardisation of semester grades in advanced-level math courses over time. Only the disadvantageous effect of a foreign country of origin reduced partly from 2007 to 2011. The ambition to achieve more comparability, equity, and fairness is not yet fulfilled.

It was assumed that this educational reform has an impact on teachers' grading practices. There are several possible explanations why an educational reform—in this case the shift from course-based to state-wide exit examinations—would not change the actions and beliefs of everyone involved in the intended way.

8.2 Changes only on the surface

Although there are different direct and indirect mechanisms of action in testing (Haertel, 2013), state-wide tests disturb teachers' actions without enhancing them (Corbett & Wilson, 1991; p. 90). Reforms and changes only scratch the surface and do not lead to a lasting improvement of practice (Cheng, 2004; Schorr & Firestone, 2004). The stability of the unintended effects of the students' background on the semester grades conveys that the external state-wide grading criteria for the examinations and the feedback processes about the accuracy of the examination grades (Maag Merki & Holmeier, 2015) were not sufficient to enhance the accuracy of the semester grades. They were constructed in the classroom as a combination of oral and written tests; the whole process laid in the teachers' hands. Teachers' attitudes towards tests and change do not automatically coincide with their actions: 'knowing about best practice and implementation are very different challenges' (Harland, McLean, Wass, Miller, & Sim, 2015, p. 539; also Cheng, 2004). Unfitting external state-wide grading criteria for the grading process in class, misunderstandings or misinterpretations of these, and inadequate teacher instructions (Black et al., 2011) may be further reasons why the higher standardisation of state-wide exit examinations and their grading had only limited 'washback' effects (Bishop, 1995; Cheng et al., 2004) on course-based semester grades.

8.3 Demanding feedback loop

With recourse to the 'standards-based accountability theory of action' (Hamilton et al., 2008) as well as to research on the responses to and analyses of test data (Monfils et al., 2004; Schorr & Firestone, 2004), it could be argued that schools and teachers did not recognise the informative value of exit examinations data for the reflection and improvement of their instruction and grading—and certainly not with regard to the course-based semester grades. It is also possible that the observed time period was too short and the experiences gained were not enough for a detailed 'feedback loop' and for effects on grading (Goldberg & Roswell, 2000; Hamilton et al., 2008; Maag Merki, 2014). For these analyses no information about teachers' experiences (individual and collective) with state-wide exit examinations were taken into consideration. It is likely that the grades were assigned by different teachers. The difficulty in this study lies in the fact that teachers have to draw conclusions from students' results in the exit examinations in one course/year to reflect about their instruction and grading and to change it in another course/year with different students. This transfer might function as an obstacle which has to be overcome. According to Altrichter, Moosbrugger, and Zuber (2016), the use of test data as a feedback instrument for school improvement requires several conditions to be fulfilled until it can influence teachers' individual actions—and this is not that simple as it might seem. However, teachers are willing to make changes (Klinger & Rogers, 2011).

8.4 Teachers' beliefs

Teachers' beliefs and values with regard to education are an important factor in realising reforms (Fives & Buehl, 2012) and may, if necessary, have to be modified initially. 'To promote educational reform then requires one to promote a different set of values and beliefs about education. [. . .] There will be no room for reforming schooling until the purposes of education are re-thought.' (Corbett & Wilson, 1991; p. 131) This change in teachers' beliefs must be accompanied by changes in policies and practices (Weinstein, 2002; p. 291). A discussion of beliefs about education, intentions, aims, and further aspects of state-wide exit examinations as well as a general understanding of these might form a breeding ground for setting changes in motion, successfully implementing reforms, and improving educational processes.

8.5 Conditions of teachers' work

The level of autonomy of several stakeholders forms another significant factor. It should be 'neither so great as to shut out external influence altogether nor so insignificant as to make educators totally

vulnerable to outside pressure' (Corbett & Wilson, 1991; p. 107). On the one hand, this offers teachers the capacity to take an active role. On the other hand, combined with processes of 'recontextualisation' (Fend, 2008), it leads to different action patterns. This is one reason why grading practices vary across teachers and schools (e.g. Bishop, 1995; Cliffordson, 2008; Gordon & Fay, 2010; Resh, 2009).

Furthermore, individual teachers and their context have to be taken into account. Teachers have to fulfil multiple tasks under permanently changing conditions which force them to set priorities. In the case of this study, standardisation and improvement of grading, especially in the context of the course-based semester grades, might not have been of prime importance from a teacher's point of view; other aspects of state-wide exit examinations may have had priority instead. Another explanation could be the teachers' attitudes to the reform: they either did not take it seriously or rejected its goals and outcomes (Terhart, 2013; cf. Corbett & Wilson, 1991).

9. LIMITATIONS AND FURTHER RESEARCH

The above analyses have several limitations. Firstly, the focus on advanced-level math courses in the German federal state of Bremen restricted the perspective and significance. The findings cannot be generalised and transferred to other subjects, levels, and samples without limitations.

Secondly, this focus entailed a reduction of the sample so that the multilevel analyses were computed with a small sample size at individual and course level. Especially with regard to the sample size at course level, one has to keep in mind that the estimates of the standard errors might by biased (Maas & Hox, 2005).

Thirdly, there was only data from one year each under the conditions of course-based (2007) and state-wide exit examinations (2011). This comparison allows indications for possible long-term effects of the implementation of state-wide exit examinations, but no conclusive answers.

Fourthly, the operationalization of ethnicity by country of origin, as well as family background by number of books, must be considered. In terms of country of origin, it is not possible to distinguish between

German-born students with foreign-born parents (2nd or 3rd generation) and German-born students with German-born parents. Furthermore, students of different countries of origin cannot be distinguished although it is well investigated that there are differences between several subgroups of students with migration background (e.g. Schräpler & Weishaupt, 2013). In addition, nothing is known about the language(s) spoken at home or with friends and how familiar the students are with German as the language of schooling. In terms of family background by number of books, no information on sociocultural processes in the families (e.g. talking about school or politics, going to the theatre; Bornkessel & Kuhnen, 2011) as well as parents' experiences with or level of education is available.

Finally, the analyses did not include other factors which could explain differences in grades, for instance students' effort or motivation and the interrelation between these factors and grades (Guillaume & Khachikian, 2011). Further studies, which take these variables into account, need to be undertaken. The same goes for the consideration of characteristics of teachers, such as stereo- types, perceptions, expectations, and assumptions (Jussim et al., 1996; van Ewijk, 2011), or schools which may have an impact on teaching and grading.

Another starting point for further analyses might be the research of Guskey (2004) who examined the stability and change in high school grades within an academic year by comparing the first and final course grades for different subgroups.

10. CONCLUSIONS

Gipps (2012) as well as Koretz (2008) advised against using only one test for important decisions so that students have several opportunities to show their achievement. That way disadvantages with a certain form of assessment could be compensated. In Germany, the combination of course-based semester grades based on achievement in several oral and written tests and the results in state-wide exit examinations for the grade point average at the end of upper secondary education, fulfils this demand. To meet the expectation that state-wide exit examinations lead to an increase in the comparability of grades, known measures to improve the quality of grades have to be examined for their suitability for

the German context (e.g. Ahmed & Pollitt, 2011; Haptonstall, 2010; Jae & Cowling, 2009; Paeplow, 2011; Wheadon & Pinot de Moira, 2013). It is particularly important to keep the teachers and their values in mind, accompany and train them so they become competent to handle grading criteria and to grade fairly (Goldberg & Roswell, 2000). This is essential in order to avoid reforming education at the expense of students; as Madaus and Clarke (2001, p. 21) said: 'The task remains of identifying strategies to achieve efficiently and effectively the desirable reform objectives—without having a negative impact on any subpopulation of students'.

REFERENCES

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*, 259–278. doi: 10.1080/0969594X.2010.546775

Altrichter, H., Moosbrugger, R., & Zuber, J. (2016). Schul- und Unterrichtsentwicklung durch Datenrückmeldung [Development of schools and instruction through data feedback]. In H. Altrichter & K. Maag Merki (Eds.), *Handbuch Neue Steuerung im Schulsystem* (2nd enl. ed., pp. 235–277). Wiesbaden: VS. doi: 10.1007/978-3-531-18942-0_9

Amengual Pizarro, M. (2010). Exploring the Washback Effects of a High-Stakes English Test on the Teaching of English in Spanish Upper Secondary Schools. *Revista Alicantina de Estudios Ingleses, 23*, 149–170. doi: 10.14198/raei.2010.23.09

Bishop, J.H. (1995). The impact of curriculum-based external examinations on school priorities and student learning. *International Journal of Educational Research*, 23, 653–752. doi: 10.1016/0883-0355(96)00001-8

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*, 451–469. doi: 10.1080/0969594X.2011.557020

Bornkessel, P., & Kuhnen, S.U. (2011). Zum Einfluss der sozialen Herkunft auf Schulleistung, Studienzuversicht und Studienintention am Ende der Sekundarstufe II [The effect of the social background on achievement, confidence, and intention for studies]. In P. Bornkessel & J. Asdonk (Eds.), *Der Übergang Schule – Hochschule. Zur Bedeutung sozialer, persönlicher und institutioneller Faktoren am Ende der Sekundarstufe II* (pp. 47–104). Wiesbaden: VS. doi: 10.1007/978-3-531-94016-8_3

Bowers, A.J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation, 17*, 141–159. doi:10.1080/13803611.2011.597112

Boykin, A.-S.M. (2010). *The relationship between high school course grades and exam scores*. E&R Report No. 09.39. Raleigh, NC: Wake County Public School System. Retrieved from http://www.wcpss.net/results/reports/2010/0939course_exams.pdf

Brookhart, S.M. (1997). A Theoretical Framework for the Role of Classroom Assessment in Motivating Student Effort and Achievement. *Applied Measurement in Education, 10*, 161–180. doi: 10.1207/s15324818ame1002_4

Bürgermeister, A. (2014). *Leistungsbeurteilung im Mathematikunterricht. Bedingungen und Effekte von Beurteilungspraxis und Beurteilungsgenauigkeit* [Performance assessment in mathematics lessons. Conditions and effects of practice and accuracy of assessment]. Münster, New York: Waxmann.

Cappellari, L., Lucifora, C., & Pozzoli, D. (2012). Determinants of grades in maths for students in economics. *Education Economics, 20*, 1–17. doi:10.1080/09645291003718340

Cheng, L. (2004). The Washback Effect of a Public Examination Change on Teachers' Perceptions Toward Their Classroom Teaching. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing. Research contexts and methods* (pp. 147–170). Mahwah, NJ: Erlbaum.

Cheng, L., & Curtis, A. (2004). Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing. Research contexts and methods* (pp. 3–17). Mahwah, NJ: Erlbaum.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing. Research contexts and methods*. Mahwah, NJ: Erlbaum.

Cliffordson, C. (2008). Differential Prediction of Study Success Across Academic Programs in the Swedish Context: The Validity of Grades and Tests as Selection Instruments for Higher Education. *Educational Assessment, 13*, 56–75. doi: 10.1080/10627190801968240

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NY: Erlbaum.

Corbett, H.D., & Wilson, B.L. (1991). *Testing, Reform, and Rebellion*. New Jersey: Ablex.

Dardanoni , V., Modica, S., & Pennisi, A. (2011). School grading and institutional contexts. *Education Economics, 19*, 475–486. doi:10.1080/09645292.2010.488482

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität* [Organizing schooling. Steering of the system, school development, and quality of instruction]. Wiesbaden: VS.

Fives, H., & Buehl, M.M. (2012). Spring cleaning for the 'messy' construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K.R. Harris, S. Graham, & T. Urdan (Eds.), *APA Educational Psychology Handbook. Volume 2: Individual Differences and Cultural and Contextual Factors* (pp. 471–499). American Psychological Association. doi: 10.1037/13274-019

Forghani-Arani, N., Geppert, C., & Katschnig, T. (2015). Wenn der Pygmalioneffekt nicht greift … [When the Pygmalioneffect fails …]. *Zeitschrift für Bildungsforschung, 5*, 21–36. doi: 10.1007/s35834-014-0104-x

Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K., & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary Educational Psychology, 41*, 1–12. doi: 10.1016/j.cedpsych.2014.10.006

Gipps, C.V. (2012). *Beyond Testing: towards a theory of educational assessment*. London & New York: Routledge.

Goldberg, G.L., & Roswell, B.S. (2000). From Perception to Practice: The Impact of Teachers' Scoring Experience on Performance-Based Instruction and Classroom Assessment. *Educational Assessment, 6*, 257–290. doi:10.1207/S15326977EA0604_3

Gomolla, M., & Radtke, F.-O. (2009). *Institutionelle Diskriminierung. Die Herstellung ethnischer Differenz in der Schule* [Institutional discrimination. Production of ethnical difference in schools] (3rd ed.). Wiesbaden: VS.

Gordon, M.E., & Fay, C.H. (2010). The Effects of Grading and Teaching Practices on Students' Perceptions of Grading Fairness. *College Teaching, 58*, 93–98. doi: 10.1080/87567550903418586

Gronlund, N.E. (1974). *Improving marking and reporting in classroom instruction*. New York: Macmillan.

Guillaume, D.W., & Khachikian, C.S. (2011). The effect of time-on-task on student grades and grade expectations. *Assessment & Evaluation in Higher Education, 36*, 251–261. doi: 10.1080/02602930903311708

Guskey, T.R. (2004, April). *Stability and Change in High School Grades*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA). San Diego, CA.

Haertel, E. (2013). How Is Testing Supposed to Improve Schooling? *Measurement: Interdisciplinary Research and Perspectives, 11*, 1–18. doi: 10.1080/15366367.2013.783752

Hamilton, L.S., Stecher, B.M., Russell, J.L., Marsh, J.A., & Miles, J. (2008). Accountability and teaching practices: school-level actions and teacher responses. In B. Fuller, M.K. Henne, & E. Hannum (Eds.), *Strong stakes, weak schools: the benefits and dilemmas of centralized accountability* (pp. 31–66). Bingley: Emerald.

Haptonstall, K.G. (2010). A*n Analysis of the Correlation between Standards-Based, Non-Standards-Based Grading Systems and Achievement as Measured by the Colorado Student Assessment Program (CSAP)*. Ann Arbor, MI: ProQuest LLC.

Hargreaves, A. (1998). The emotional practice of teaching. *Teaching and Teacher Education, 14*, 835–854. doi: 10.1016/S0742-051X(98)00025-0

Harland, T., McLean, A., Wass, R., Miller, E., & Sim, K.N. (2015). An assessment arms race and its fallout: high-stakes grading and the case for slow scholarship. *Assessment & Evaluation in Higher Education, 40*, 528–541. doi: 10.1080/02602938.2014.931927

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf Schüler- und Schulklassenebene in Primar- und Sekundarstufe* [What do grades in math measure? Correlates of grades in Math on student- and class-level in primary and secondary schools]. Münster: Waxmann.

Hofer, S.I. (2015). Studying Gender Bias in Physics Grading: The role of teaching experience and country. *International Journal of Science Education, 37*, 2879–2905. doi: 10.1080/09500693.2015.1114190.

Holmeier, M. (2013). *Leistungsbeurteilung im Zentralabitur* [Performance assessment in state-wide A-level exams]. Wiesbaden: VS. doi: 10.1007/978-3-531-19725-8

Jae, H., & Cowling, J. (2009). Objectivity in Grading: The Promise of Bar Codes. *College Teaching, 57*, 51–55. doi: 10.3200/CTCH.57.1.51-55

Jussim, L., Eccles, J., & Madon, S. (1996). Social Perception, Social Stereotypes, and Teacher Expectations: Accuracy and the Quest for the Powerful Self-Fulfilling Prophecy. *Advances in Experimental Social Psychology, 28*, 281–388. doi: 10.1016/S0065-2601(08)60240-3

Kahnert, J. (2014). *Das Zentralabitur im Fach Mathematik. Eine empirische Analyse von Abitur- und TIMSS-Daten im Vergleich* [A-level examinations in Math. A comparable empirical analysis of exam- and TIMSS-data]. Münster & New York: Waxmann.

Klapp Lekholm, A., & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation, 15*, 1–23. doi:10.1080/13803610802470425

Klein, E.D., & van Ackeren, I. (2011). Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels. *Studies in Educational Evaluation, 37*, 180–188. doi: 10.1016/j.stueduc.2012.01.002

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzen und Unterrichtsschwerpunkte [Achievement in Advanced Mathematics and Physics: Theoretical background, competence levels and curricular priorities]. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 57–128). Opladen: Leske + Budrich.

Klieme, E. (2003). Benotungsmaßstäbe an Schulen: Pädagogische Praxis und institutionelle Bedingungen. Eine empirische Analyse auf der Basis der PISA-Studie [Standards for grading in schools: Educational practice and institutional conditions. An empirical analysis on the basis of the PISA-study]. In H. Döbert, B. von Kopp, R. Martini, & M. Weiß (Eds.), *Bildung vor neuen Herausforderungen. Historische Bezüge – Rechtliche Aspekte – Steuerungsfragen – Internationale Perspektiven* (pp. 195–210). Neuwied: Luchterhand.

Klieme, E., Bürgermeister, A., Harks, B., Blum, W., Leiß, D., & Rakoczy, K. (2010). Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht. Projekt Co2CA [Performance assessment and competence models in classrooms in Math. The project Co2CA]. In E. Klieme, D. Leutner, & M. Kenk (Eds.), Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes (Special Issue 56). *Zeitschrift für Pädagogik, 64–74.*

Klinger, D.A., & Rogers, W.T. (2011). Teachers' Perceptions of Large-Scale Assessment Programs Within Low-Stakes Accountability Frameworks. *International Journal of Testing, 11*, 122–143. doi: 10.1080/15305058.2011.552748

Koretz, D. (2008). *Measuring Up. What Educational Testing Really Tells Us*. Cambridge et al.: Harvard University Press.

Kühn, S.M. (2011). Exploring the use of statewide exit exams to spread innovations – The sample of Context in science tasks from an international comparative perspective. *Studies in Educational Evaluation, 37*, 189–195. doi: 10.1016/j.stueduc.2012.01.003

Lazarides, R., & Watt, H.M.G. (2015). Girls' and boys' perceived mathematics teacher beliefs, classroom learning environments and mathematical career intentions. *Contemporary Educational Psychology, 41*, 51–61. doi: 10.1016/j.cedpsych.2014.11.005

Maag Merki, K. (Ed.). (2012). *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* [State-wide A-level exams. The longitudinal analysis of processes and effects of the implementation of state-wide A-level exams in two German federal states]. Wiesbaden: VS. doi: 10.1007/978-3-531-94023-6

Maag Merki, K. (2014). Das quasi-experimentelle Design in der Educational Governance-Forschung? Herausforderungen, Möglichkeiten und Grenzen am Beispiel der Analyse der Wirksamkeit der Einführung zentraler Abiturprüfungen [The quasi-experimental design in research on Educational Governance? Challenges, possibilities, and limitations through the example of the analysis of the effectiveness of the implementation of state-wide A-level exams]. In K. Maag Merki, R. Langer, & H. Altrichter (Eds.), E*ducational Governance als Forschungsperspektive. Strategien. Methoden. Ansätze* (2nd ed., pp. 51–83). Wiesbaden: Springer VS. doi: 10.1007/978-3-531-19148-5_2

Maag Merki, K., & Holmeier, M. (2015). Comparability of semester and exit exam grades: long-term effect of the implementation of state-wide exit exams. *School Effectiveness and School Improvement, 26,* 57–74. doi: 10.1080/09243453.2013.861353

Maas, C.J.M., & Hox, J.J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology, 1*, 86–92. doi: 10.1027/1614-1881.1.3.86

Maaz, K., Baeriswyl, F., & Trautwein, U. (2011). *Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Eine Studie im Auftrag der Vodafone Stiftung Deutschland* [Censored origin? Performance assessment and social inequalities in schools. A study by order of Vodafone Foundation Germany]. Retrieved from https://www.vodafone-stiftung.de/alle_publikationen.html?&tx_newsjson_pi1%5BshowUid%5D=44&cHash=77c9bde76786dbdff61bdb720d3dff8c

Madaus, G.F., & Clarke, M. (2001). *The Adverse Impact of High Stakes Testing on Minority Students: Evidence from 100 Years of Test Data*. Retrieved from http://files.eric.ed.gov/fulltext/ED450183.pdf

Marsh, H.W., & O'Mara, A.J. (2010). Long-Term Total Negative Effects of School-Average Ability on Diverse Educational Outcomes. Direct and Indirect Effects of the Big-Fish-Little-Pond Effect. *Zeitschrift für Pädagogische Psychologie, 24*, 51–72. doi: 10.1024/1010-0652/a000004

Maué, E. (2013). Vergleichbarkeit von Abiturnoten—eine Fiktion? Längerfristige Effekte der Implementation zentraler Abiturprüfungen in Bremen [Comparability of exit exam grades—a fiction? Long-term effects of the implementation of state-wide exit exams in upper secondary education in Bremen]. In S. U. Kuhnen (Ed.), *Von der Schule zur Hochschule. Analysen, Konzeptionen und Gestaltungsperspektiven des Übergangs* (pp. 114–128).Münster: Waxmann.

Monfils, L.F., Firestone, W.A., Hicks, J.E., Martinez, M.C., Schorr, R.Y., & Camilli, G. (2004). Teaching to the Test. In W.A. Firestone, R.Y. Schorr, & L.F. Monfils (Eds.), *The Ambiguity of Teaching to the Test. Standards, Assessment, and Educational Reform* (pp. 37–61). Mahwah, NJ & London: Erlbaum.

Møen, J., & Tjelta, M. (2010). Grading Standards, Student Ability and Errors in College Admission. *Scandinavian Journal of Educational Research, 54*, 221–237. doi:10.1080/00313831003764503

Neumann, M., Trautwein, U., & Nagy, G. (2011). Do state-wide examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. *Studies in Educational Evaluation, 37*, 206–217. doi: 10.1016/j.stueduc.2012.02.002

Neumann, M., Nagy, G., Trautwein, U., & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen. Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen [Comparability of high-school diploma: Differences in achievement and grading between students in Hamburg and Baden-Württemberg and the central role of state-wide exit examinations]. *Zeitschrift für Erziehungswissenschaft, 12*, 691–714. doi: 10.1007/s11618-009-0099-6

OECD (2012). *Grade Expectations: How Marks and Education Policies Shape Students' Ambitions.* PISA OECD Publishing. doi: 10.1787/9789264187528-en

Paeplow, C.G. (2008). *Middle School Grading: Wake County Public School System (WCPSS) 2006-07 and 2007-08.* E&R Report No. 08.16. Raleigh, NC: Wake County Public School System. Retrieved from http://www.wcpss.net/results/reports/2008/0816ms_grading2008.pdf

Paeplow, C.G. (2011). *Easy as 1, 2, 3: Exploring the Implementation of Standards-Based Grading in Wake County Elementary Schools.* Dissertation at North Carolina State University. Retrieved from http://repository.lib.ncsu.edu/ir/bitstream/1840.16/7242/1/etd.pdf

Prodromou, L. (1995). The backwash effect: from testing to taching. *ELT Journal, 49*, 13–25. doi: 10.1093/elt/49.1.13

Rakoczy, K., Klieme, E., Bürgermeister, A., & Harks, B. (2008). The Interplay Between Student Evaluation and Instruction. Grading and Feedback in Mathematics Classrooms. *Zeitschrift für Psychologie, 216*, 111–124. doi: 10.1027/0044-3409.216.2.111

Raudenbush, S.W., Bryk, A.S., & Congdon, R. (2004). *HLM 6 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.

Reardon, S.F., Atteberry, A., Arshan, N., & Kurlaender, M. (2009). *Effects of the California High School Exit Exam on Student Persistence, Achievement, and Graduation*. Working Paper 2009-12. Stanford University, Institute for Research on Education Policy & Practice. Retrieved from http://web.stanford.edu/group/cepa/workingpapers/WORKING_PAPER_2009_12.pdf

Resh, N. (2009). Justice in grades allocation: teachers' perspective. *Social Psychology of Education, 12*, 315–325. doi: 10.1007/s11218-008-9073-z

Resh, N. (2010). Sense of justice about grades in school: is it stratified like academic achievement? *Social Psychology of Education, 13*, 313–329. doi: 10.1007/s11218-010-9117-z

Rosenthal, R., & Jacobson, L. (1992). P*ygmalion in the Classroom. Teacher Expectation and Pupils' Intellectual Development* (enl. ed.). New York: Irvington.

Rubie-Davis, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology, 76*, 429–444. doi: 10.1348/000709905X53589

Schorr, R.Y., & Firestone, W.A. (2004). Conlusion. In W. A. Firestone, R. Y. Schorr, & L. F. Monfils (Eds.), *The Ambiguity of Teaching to the Test. Standards, Assessment, and Educational Reform* (pp. 159–168). Mahwah, NJ: Erlbaum.

Schräpler, J.-P., & Weishaupt, H. (2013). Auswirkungen des Zentralabiturs auf den Abiturerfolg an Gymnasien und Gesamtschulen in Nordrhein-Westfalen [Effects of state-wide exit exams on graduation at academic-track secondary schools and comprehensive schools in North Rhine-Westphalia]. In N. McElvany & H. G. Holtappels (Eds.), *Empirische Bildungsforschung. Theorien, Methoden, Befunde und Perspektiven* (pp. 249–266). Münster et al.: Waxmann.

Stevens, P.A.J., & Görgöz, R. (2010). Exploring the importance of institutional contexts for the development of ethnic stereotypes: a comparison of schools in Belgium and England. *Ethnic and Racial Studies, 33*, 1350–1371. doi: 10.1080/01419870903219243

Tenenbaum, H.R., & Ruck, M.D. (2007). Are Teachers' Expectations Different for Racial Minority Than for European American Students? A Meta-Analysis. *Journal of Educational Psychology, 99*, 253–273. doi: 10.1037/0022-0663.99.2.253

Terhart, E. (2013). Teacher resistance against school reform: reflecting an inconvenient truth. *School Leadership & Management, 33*, 486–500. doi: 10.1080/13632434.2013.793494

Thorsen, C. (2012). Dimensions of Norm-Referenced Compulsory School Grades and their Relative Importance for the Prediction of Upper Secondary School Grades. *Scandinavian Journal of Educational Research*. iFirst Article, 1–20. doi: 10.1080/00313831.2012.705322

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation, 18*, 153–172. doi:10.1080/13803611.2012.659929

Tierney, R.D., Simon, M., & Charland, J. (2011). Being Fair: Teachers' Interpretation of Principles for Standards-Based Grading. *The Educational Forum, 75*, 210–227. doi: 10.1080/00131725.2011.577669

Trouilloud, D., Sarrazin, P., Martinek, T., & Guillet, E. (2002). The Influence of Teacher Expectations on Students Achievement in Physical Education Classes: Pygmalion Revisited. *European Journal of Social Psychology, 3*2, 591–607. doi: 10.1002/ejsp.109

van Ackeren, I., Block, R., Klein, E.D., & Kühn, S.M. (2012). The Impact of State-Wide Exit Exams in Germany: A Descriptive Case Study of Three German States with Differing Low Stakes Exam Regimes. *Education Policy Analysis Archives, 20*, 1–28. Retrieved from http://epaa.asu.edu/ojs/article/view/1011/964

van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review, 30*, 1045–1058. doi: 10.1016/j.econedurev.2011.05.008

Verordnung über die Gymnasiale Oberstufe (GyO-VO) vom 1. August 2005 [Regulation on the academic-track upper school (GyO-VO) from August 1st, 2005]. Retrieved from http://transparenz.bremen.de/sixcms/detail.php?gsid=bremen2014_tp.c.67089.de&template=00_html_to_pdf_d

Watanabe, Y. (2004). Methodology in Washback Studies. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing. Research contexts and methods* (pp. 19–36). Mahwah, NJ: Erlbaum.

Weinstein, R. S. (2002). *Reaching Higher. The Power of Expectations in Schooling*. Cambridge, MA & London: Harvard University Press.

Wheadon, C., & Pinot de Moira, A. (2013). Gains in marking reliability from item-level marking: is the sum of the parts better than the whole? *Educational Research and Evaluation, 19*, 665–679. doi: 10.1080/13803611.2013.828629

Woessmann, L., Luedemann, E., Schuetz, G., & West, M.R. (2009). *School Accountability, Autonomy and Choice around the World*. Cheltenham & Northhampton, MA: Edward Elgar.

Publikation 3: Emotionales Erleben des Zentralabiturs von Lehrpersonen

Maué, E., Maag Merki, K. & Oerke, B. (2012). Emotionales Erleben des Zentralabiturs von Lehrpersonen in Bremen. Längerfristige Effekte der Implementation zentraler Abiturprüfungen. In S. Hornberg & M. Parreira do Amaral (Hrsg.). Deregulierung im Bildungswesen (S. 109-130). Münster et al.: Waxmann.

1. EINLEITUNG

Im Jahr 2018 legen Schülerinnen und Schüler in Deutschland erstmals bundesweit in den Fächern Deutsch, Englisch und Mathematik einheitliche Abiturprüfungen ab – zumindest, wenn die Verantwortlichen in der Bildungspolitik die Empfehlungen des Aktionsrats Bildung zum Gemeinsamen Kernabitur umsetzen (vbw – Vereinigung der Bayerischen Wirtschaft e. V. , 2011)[1]. Der Aktionsrat Bildung schreibt dem Gemeinsamen Kernabitur durch eine „höhere Vergleichbarkeit der Abiturleistungen positive Rückwirkungen auf die Qualität der schulischen Leistungen" sowie eine Veränderung der „Rolle der Lehrkräfte von ‚Richtern' zu ‚Coaches', die die Schülerinnen und Schüler auf die extern gesetzten Anforderungen vorbereiten" (ebd., S. 17), zu. Auch wenn in den letzten Jahren durch die bis auf Rheinland-Pfalz mittlerweile flächendeckend eingeführten zentralen Abiturprüfungen Tendenzen einheitlicher Standards zu erkennen sind, bestehen weiterhin deutliche bundeslandspezifische Unterschiede bei deren Anlage, Durchführung und Bewertung (ebd.). Somit wäre die Implementation eines Gemeinsamen Kernabiturs eine große Veränderung der föderalistisch geprägten Bildungslandschaft und hätte in vielerlei Hinsicht Auswirkungen auf alle Beteiligten des Mehrebenensystems Schule (Ditton, 2007; Fend, 2008). Einen entscheidenden Aspekt bei der Umsetzung von Reformen stellen laut Hargreaves (2004) die Emotionen von Lehrpersonen dar – Emotionen sind „at the heart of teaching" (1998, S. 835). Aus diesem Grund nimmt der vorliegende Beitrag die Entwicklung des emotionalen Erlebens in den ersten Jahren nach der Einführung zentraler Abiturprüfungen im Jahr 2007 bei Bremer Lehrpersonen in den Blick.

Zunächst werden grundlegende Begriffe definiert sowie der Forschungsstand und sich daraus ableitende Fragestellungen und Hypothesen vorgestellt. Eine kurze Beschreibung von Design, Methodik und

[1] im Folgenden vbw

Stichprobe der Studie „Implementation und Auswirkungen neuer Steuerungsstrukturen im Schulwesen am Beispiel zentraler Abiturprüfungen" bildet die Grundlage für die Präsentation und Diskussion ausgewählter quer- und längsschnittlicher Ergebnisse zu den Emotionen Bremer Lehrpersonen bezüglich des Zentralabiturs. Ein Ausblick rundet den Beitrag ab.

2. BEGRIFFLICHE KLÄRUNGEN

Vielfach wird der Lehrberuf als besonders stressig und psychisch belastend bezeichnet (z.B. Krause, Dorsemagen & Alexander, 2011; Rothland, 2007a). Laut Stöckli (1992, S. 4) „zählt die Schule zu den ausgeprägtesten Stress-Ökologien". Entsprechend häufig werden Belastung, Beanspruchung und Stress von Lehrpersonen untersucht. Rudow (1990, 1994) entwickelte ein Rahmenmodell zur Analyse der Belastung und Beanspruchung im Lehrberuf, welches Böhm-Kasper (2004) so modifizierte, dass es sowohl für die Ebene der Lehrpersonen als auch für die der Schülerinnen und Schüler gilt. Rudow (1994) differenziert die Mehrfachbelastung des Lehrberufs in objektive, subjektive bzw. psychische und kognitive Belastung aus. Als *objektive Belastung* gelten die allgemeinen und möglicherweise Beanspruchung hervorrufenden pädagogischen Arbeitsaufgaben und -bedingungen (wertneutral), welche durch die Wahrnehmung, Bewertung und kognitive Verarbeitung in *subjektive* bzw. *psychische*, aber auch in positive oder negative *emotionale Belastungen* transformiert werden. Diese können einerseits individueller, andererseits kollektiver Natur sein (ebd.). Die Diskrepanz des Vergleichs „von Bedürfnissen oder Motiven und deren wahrgenommenen bzw. antizipierten Realisierungsmöglichkeiten […] bestimmt letztendlich die Qualität der emotionalen Belastung" (ebd., S. 42). Die *kognitive Belastung* ist von den bei Herausforderungen eingesetzten kognitiven Ressourcen abhängig. Die „zeitlich unmittelbare Konfrontation der psychischen und physischen Handlungsvoraussetzungen des Individuums mit den Tätigkeitsanforderungen" (ebd., S. 45) kennzeichnet die *Beanspruchung*. Positiv ist diese, wenn Anforderungen als Herausforderung bewertet werden, was die pädagogische Handlungskompetenz fördert, negativ, wenn sie als Bedrohung erscheinen, damit die pädagogische Handlungskompetenz einschränken und zu Stress führen können (ebd.; Sieland, 2007). Die Sichtweise von und der Umgang mit Beanspruchung wirken sich also stets auf die weitere pädagogische Tätigkeit und deren Erleben aus. Die Auseinandersetzung des Individuums mit der Umwelt ist auch für *Stress* bedeutsam. „Stress is

any event in which environmental or internal demands tax or exceed the adaptive resources of an individual, social system, or tissue system" (Lazarus & Launier, 1978, S. 296, zitiert nach Schwarzer, 2000, S. 11). Stress basiert damit auf einer wechselseitigen Beziehung zwischen der Wahrnehmung einer Situation und den individuellen Ressourcen oder Einstellungen (Buchwald, 2011). Beanspruchungs- und stressauslösende Faktoren können sowohl äußere als auch innere Einflüsse sein (Stähling, 1998; siehe dazu ebenfalls Böhm-Kasper, 2004).

Mit *zentralen (Abschluss-)Prüfungen* werden Begriffe wie Standardisierung, Vergleichbarkeit, Fairness und Qualitätssicherung sowie verschiedene Ziele verbunden: Einheitliche Prüfungen sollen zu einer größeren Standardisierung führen und durch die Sicherung von (Mindest-)Standards das Leistungsniveau generell erhöhen. Zudem ermöglichen zentrale (Abschluss-)Prüfungen eine bessere Vergleichbarkeit der Leistungen. Die Benotung soll objektiver und damit gerechter ausfallen, da sie anhand verbindlicher externer Kriterien erfolgt, was zudem die diagnostische Kompetenz der Lehrkräfte fördern soll. Eine breitere Themenabdeckung sowie innovative Curricula und Aufgabenformate sollen mit mehr Qualität des Unterrichts einhergehen. Darüber hinaus wird eine Steigerung der Leistungsbereitschaft durch die Erhöhung der extrinsischen Motivation erwartet (Block, Klein, van Ackeren & Kühn, 2011). Obwohl zwischen den deutschen Bundesländern Unterschiede bezüglich der Konzeption zentraler Abiturprüfungen bestehen – für einen Überblick über bundeslandübergreifende und -spezifische Regelungen siehe vbw (2011) und Kühn (2012) – lassen sich zentrale Merkmale festhalten. Im Gegensatz zu dezentralen Abiturprüfungen erstellen nicht die jeweiligen Kurslehrpersonen die Prüfungsaufgaben, sondern von einer externen Aufgabenkommission bzw. der obersten Schulaufsichtsbehörde beauftragte Lehrkräfte, deren Vorschläge anschließend ausgewählt werden (Kühn, 2012). Beiden Formen gemeinsam ist die dezentrale Korrektur durch die Kurslehrpersonen sowie die Zweitkorrektur durch andere Fachlehrpersonen, meistens derselben Schule.

3. FORSCHUNGSSTAND

Die Forschung zu *schulischer Belastung, Beanspruchung und Stress* stützt sich vorrangig auf modifizierte allgemeine Modelle zu Belastung, Beanspruchung und Stress. Eine wichtige Rolle spielen dabei das

Rahmenmodell der Belastung und Beanspruchung von Rudow und das transaktionelle Stressmodell von Lazarus. Einen Überblick über diese sowie weitere relevante Modelle bieten z.B. Böhm-Kasper (2004), Herzog (2007), Krause et al. (2011), Soltau und Mienert (2010) sowie van Dick und Stegmann (2007). Vielfältige Faktoren und Situationen können als sogenannte Stressoren fungieren und Beanspruchung oder Stress hervorrufen. Eine Zusammenfassung von Stressoren, aber auch positiven äußeren Ressourcen, bieten u.a. Rudow (1994) und Krause et al. (2011). Zur Erfassung von Beanspruchung und Stress dient das Fragebogenverfahren „Arbeitsbezogenes Verhaltens- und Erlebensmuster" (AVEM) von Schaarschmidt und Fischer, das mittels der Merkmale Arbeitsengagement, Widerstandskraft gegenüber Belastungen und berufsbegleitende Emotionen die vier Beanspruchungsmuster „Gesundheitstyp", „Schontyp", „Risikotyp A" und „Risikotyp B" unterscheidet (Schaarschmidt & Kieschke, 2007; siehe auch Klusmann, Kunter & Trautwein, 2009). Im Mittelpunkt der „Potsdamer Lehrerstudie" steht das mit dem AVEM erfasste Verhalten von Lehrpersonen in Anbetracht der beruflichen Anforderungen und damit verbundene Gesundheitsressourcen und -risiken (Schaarschmidt & Kieschke, 2007). Als Ergebnis ist festzuhalten, dass der Frauenanteil in den Risikotypen größer ausfällt und eine „progressive Verschlechterung der Beanspruchungssituation über die Berufsjahre" stattfindet, ebenfalls in stärkerem Maß bei den Frauen (ebd., S. 90). Bieri (2006) untersucht verschiedene Facetten beruflicher Zufriedenheit und Belastungen von Aargauer Lehrpersonen. Eine Besonderheit liegt in der Stichprobe – neben im Schuldienst tätigen Lehrkräften werden auch solche, die ihre Stelle kündigten, erfasst.

Förderlich für den Umgang mit Beanspruchung und Stress ist eine hohe *Selbstwirksamkeitserwartung*, bezogen auf das schulische Umfeld, also die „Überzeugung der Lehrkräfte […], dass sie auch in Anbetracht von Schwierigkeiten in der Lage sind, verschiedene Anforderungen wie den Umgang mit schwierigen Schülern, die Elternkontakte oder die Unterrichtsgestaltung adäquat zu bewältigen" (Klusmann et al., 2009, S. 202). Eine hohe Selbstwirksamkeitserwartung wirkt sich sowohl bei Schülerinnen und Schülern als auch bei Lehrpersonen förderlich auf die Motivation und Leistungen, aber auch auf das Wohlbefinden und die (Berufs-)Zufriedenheit aus. Neben der individuellen ist auch die kollektive Selbstwirksamkeit von Bedeutung, also die Gewissheit, aufgrund der individuellen Ressourcen der Gruppenmitglieder Herausforderungen gemeinsam zu bestehen. Diese Überzeugung beeinflusst beispielsweise die Umsetzung von Reformen und Innovationen im Unterricht positiv (Schwarzer & Warner, 2011).

Für das schulische Umfeld von Bedeutung sind Auswirkungen von Beanspruchung und Stress auf die *Arbeitszufriedenheit* von Lehrpersonen. Merz definiert Arbeitszufriedenheit als das „Ergebnis eines Vergleichs von Merkmalen des Berufes bzw. der Berufssituation und den subjektiven Erwartungen und Bedürfnissen des Berufstätigen" (1979, S. 59, zitiert nach Gehrmann, 2007). Dabei spielen einerseits kognitive und emotionale Bewertungen eine Rolle, andererseits wird Arbeitszufriedenheit weitergehend als emotionaler Zustand interpretiert (siehe dazu Rudow, 1994). Lehrkräfte erleben ihre Aufgaben vorrangig als bewältigbar und weniger als Belastung, was – alters- und geschlechtsunspezifisch – zu einer Zufriedenheit mit ihrem Beruf führt (Gehrmann, 2007). So eindeutig scheint die Befundlage jedoch nicht zu sein, da Gehrmann (2007) auch von geschlechtsspezifischen Differenzen bei der Zufriedenheit mit dem Beruf berichtet, wonach die Lehrerinnen unzufriedener seien. Bieri (2006) sowie Jäger (2012b) kommen hingegen zu dem Ergebnis, dass Lehrerinnen zufriedener mit ihrem Beruf sind als Lehrer.

Einen positiven Effekt auf die Bewältigung von Belastung und Stress haben ein gutes *Schulklima* sowie eine unterstützende Haltung des Kollegiums und der Schulleitung (Gehrmann, 2007; Rudow, 1994; Schaarschmidt & Kieschke, 2007). In Bezug auf die Auswirkungen von *Kooperation* im Kollegium sind die Forschungsbefunde nicht eindeutig, u.a. durch unterschiedliche Definitionen und Erhebungen von Kooperation bedingt (Fussangel, 2008; Kamski, 2011). Einerseits kann Kooperation als Stressor wirken, da sie zusätzliche Absprachen und damit Aufwand erfordert sowie konträr zum Autonomiebedürfnis der Lehrpersonen steht (z.B. Fussangel, Dizinger, Böhm-Kasper & Gräsel, 2010; Soltau & Mienert, 2009), andererseits kann sie Unterstützung und Entlastung bieten sowie Unsicherheit reduzieren (Fussangel & Gräsel, 2011; Soltau & Mienert, 2010). Zudem steht Kooperation in einem Zusammenhang mit der Selbstwirksamkeit: Je höher diese ausgeprägt ist, desto häufiger kooperieren Lehrpersonen (Fussangel et al., 2010), was wiederum zu einem guten Schulklima, einer höheren Arbeitszufriedenheit und einer größeren Bereitschaft, Reformen umzusetzen, führt (Fussangel & Gräsel, 2011; Rothland, 2007b).

Bei der Umsetzung von Reformen spielen *Emotionen* eine große Rolle, da Veränderungen und Emotionen untrennbar miteinander verbunden sind: „There is no human change without emotion and there is no emotion that does not embody a momentary or monomentous process of change" (Hargreaves, 2004, S. 287). Für den vorliegenden Beitrag sind primär solche Emotionen von Interesse, die mit Veränderungen, insbesondere Reformen im schulischen Umfeld, einhergehen, z.B. *Unsicherheit*. Sie ist

ein „konstitutives Element des konkreten alltäglichen LehrerInnenhandelns" (Lüsebrink, 2002, S. 44). Munthe (2001, S. 171) unterscheidet drei Aspekte von „teacher certainty": „Didactic certainty can be defined as ‚certainty about reaching all students academically', practical certainty as ‚certainty about methods used to teach and intervene' and relational certainty as ‚certainty about relationship building among students and parents'". Unsicherheit resultiert beispielsweise aus mangelndem „standardisiertem beruflichem Wissen" (Soltau & Mienert, 2010, S. 763), mangelnder Kontrolle der (längerfristigen) Ergebnisse und Wirkungen, ungewissem Anteil an (Miss-)Erfolgen der Schülerinnen und Schüler, unklaren Zielen, widersprüchlichen Rollenerwartungen, der nötigen Balancefindung zwischen Bedürfnissen der Individuen und der Klasse, Handlungs- und Zeitdruck sowie wenig institutionalisiertem Feedback (ebd.; Floden & Buchmann, 1993; Lortie, 1975; Lüsebrink, 2002; siehe auch die Ungewissheitsantinomie von Helsper, 2000). Einige dieser vielfältigen Faktoren von Unsicherheit lassen sich mit (Berufs-)Erfahrung zumindest verringern. „Novizen haben *ein* zentrales Problem, die Gewinnung von Handlungssicherheit; sie wollen wissen, wie sie erfolgreich unterrichten können" (Oelkers, 2000, S. 132; Hervorhebung im Original). Diese Handlungssicherheit werden sie mit zunehmender Erfahrung erlangen (Lortie, 1975) und damit unsichere Situationen erfolgreicher bewältigen können. Erfahrenere Lehrkräfte nehmen Unsicherheit als Bestandteil ihrer Rolle wahr und nicht als Ausdruck persönlicher Unzulänglichkeit (Lange & Burroughs-Lange, 1994). „Teachers can reduce some uncertainties by deepening and strengthening their pedagogical knowledge and skills during initial preparation or later in their careers" (Floden & Buchmann, 1993, S. 379).

Hingegen scheint es, als könnten jüngere Lehrpersonen besser mit Reformen des Schulsystems umgehen als ältere Lehrkräfte. So kommt Hargreaves in einer Untersuchung zu dem Ergebnis, dass für junge Lehrkräfte Unsicherheit nicht nur zum beruflichen Umfeld, sondern zum Leben generell dazugehört. Dies führt zu einem anderen Umgang mit Unsicherheit, weshalb sie Reformen gegenüber offener eingestellt sind. „This new generation of teachers is also more flexible, adaptable, accepting and even enthusiastic in its dealings with educational and other kinds of change" (2005, S. 972). Hargreaves zeigt weiterhin, dass Emotionen gegenüber Reformen bei Lehrpersonen von (Dienst-)Alter, Karrierestatus, Generation, fachspezifischen und persönlichen Orientierungen abhängen. Demnach sind ältere Lehrkräfte Reformen gegenüber eher negativ eingestellt, während jüngere Lehrpersonen sowie Lehrerinnen positive Haltungen aufweisen (Hargreaves, 2004). Reformen können als Stressoren wirken (Krause et al., 2011)

und mit eher negativen Emotionen bei Lehrpersonen einhergehen, v.a. wenn diese starken Einfluss auf den Unterricht haben (Brown, Ralph & Brember, 2002; Hargreaves, 2004; Kelchtermans, 2005; Munthe, 2003). Berufsbezogene Überzeugungen dienen dabei als Filter und beeinflussen die Haltung gegenüber Reformen (Krapp & Hascher, 2011; Reusser, Pauli & Elmer, 2011).

Eine wesentliche bildungspolitische Reform des letzten Jahrzehnts war in einigen deutschen Bundesländern die Abkehr von dezentralen hin zu zentralen Abiturprüfungen. Die mit zentralen (Abschluss-) Prüfungen verbundenen Ziele (siehe Kapitel 2) gehen jedoch mit negativen Folgen für die Beteiligten einher, vor allem, wenn es sich um high-stakes testing handelt (Amrein & Berliner, 2002). So erleben Schülerinnen und Schüler in Ländern mit zentralen Abschlussprüfungen mehr Stress, Angst und Müdigkeit (Nichols & Berliner, 2007; Pedulla, Abrams, Madaus, Russell, Ramos & Miao, 2003; Ryan, Ryan, Arbuthnot & Samuels, 2007). Ähnliches gilt für Lehrpersonen: Sie sind mit ihrer Arbeit weniger zufrieden und empfinden größeren Stress. „Teachers in CBEEES nations perceived themselves to have lower relative status and were significantly more likely to report wanting to leave the profession if an opportunity came along. […] teachers work harder and are under much greater stress because their ‚success' or ‚lack of success' as teachers is now more visible to others" (Bishop, 1999, S. 389f.). Darüber hinaus besteht die Gefahr einer Einengung der im Unterricht behandelten Themen sowie von zu wenig Zeit für aktuelle und lokale Themen oder einzelschulische Bedingungen (Block et al., 2011; Jäger, 2012a).

Andererseits können zentrale (Abschluss-)Prüfungen auch (psychisch) entlastende Wirkungen haben, da sie beispielsweise die Lehrkräfte von der zeit- und arbeitsintensiven Erstellung der Prüfungsfragen entbinden (Böhm-Kasper & Weishaupt, 2002; Maag Merki, 2008). Durch die Verlagerung eines Teiles der Prüfungen nach „außen" wandelt sich darüber hinaus die Rolle der Lehrpersonen hin zu Begleiterinnen und Begleitern bzw. Verbündeten der Schülerinnen und Schüler und damit das Verhältnis zwischen diesen Personengruppen. Dies verändert zudem die Unterrichtsgestaltung (Maag Merki, 2008; vdw, 2011).

Mit welchen Effekten die Einführung zentraler Abiturprüfungen in Bremen und Hessen im Jahr 2007 sowohl für Lehrpersonen als auch Schülerinnen und Schülern einhergeht, untersuchen Maag Merki und Mitarbeiterinnen über einen Zeitraum von mehreren Jahren (Maag Merki, 2012). Oerke (2012a)

fokussiert dabei beispielsweise auf das emotionale Erleben von Schülerinnen, Schülern und Lehrkräften in den Jahren 2007 bis 2009.

Als Forschungsdesiderat ist festzuhalten, dass die Erforschung des emotionalen Erlebens von Lehrpersonen gegenüber Reformen vorrangig auf Querschnittanalysen beruht. Die Veränderung über die Zeit nach der Einführung zentraler Abiturprüfungen ist bisher lediglich in einer 3-Jahres-Perspektive analysiert worden. Hier setzt der vorliegende Beitrag durch den Einbezug des Zeitraumes von fünf Jahren an.

4. FRAGESTELLUNG UND HYPOTHESEN

Der Beitrag fokussiert auf das emotionale Erleben von Bremer Lehrpersonen in Zusammenhang mit der Einführung zentraler Abiturprüfungen im Jahr 2007 und dessen Entwicklung bis zum Jahr 2011. Das emotionale Erleben wird mittels der Dimension „Unsicherheit" sowie der mit dem Erleben von Belastung, Beanspruchung und Stress verbundenen Faktoren „Leistungsdruck", „Entlastung" und „Arbeitsunzufriedenheit" operationalisiert. Folgende Forschungsfragen liegen ihm zugrunde: Wie empfinden Lehrpersonen in Bremen das Zentralabitur fünf Jahre nach der Implementation? Wie entwickelt sich das emotionale Erleben des Zentralabiturs bei Lehrpersonen in Bremen seit der Einführung im Jahr 2007? Welche Prädiktoren erklären im Jahr 2011 – dem letzten Jahr der vorliegenden längsschnittlichen Erhebung – die Facetten Unsicherheit, Leistungsdruck, Entlastung und Arbeitsunzufriedenheit?

Entsprechend den Forschungsbefunden (z.B. Oerke, 2012a; Lange & Burroughs-Lange, 1994) ist davon auszugehen, dass sich die Unsicherheit gegenüber dem Zentralabitur mit zunehmender individueller und kollektiver Erfahrung bzw. der seit der Einführung vergangenen Zeit verringert, dass also ein Unterschied zwischen den Erhebungen 2007 und 2011 besteht (Hypothese 1). Auch wenn Oerke (2012a) lediglich eine geringe, aber kontinuierliche Abnahme des Leistungsdrucks in Bremen zwischen den Jahren 2007 und 2009 sowie einen kleinen Effekt der Erfahrung mit den zentralen Abiturprüfungen auf den empfundenen Leistungsdruck feststellt, wird angenommen, dass sich dieser Trend weiter fortsetzt und der Leistungsdruck im Jahr 2011 ebenfalls geringer ausfällt als 2007 (Hypothese 2). Da beispielsweise die zeitintensive Erstellung der Prüfungsaufgaben durch zentrale Abiturprüfungen wegfällt, sollte

– in Analogie zu den Ergebnissen von Oerke (2012a) – zunehmend eine entlastende Komponente des Zentralabiturs wahrgenommen werden (Hypothese 3). Laut Gehrmann (2007) ist die Arbeitszufriedenheit von Lehrpersonen zeitlich stabil, was der Vergleich der Daten der Jahre 2007 und 2009 bestätigt (Jäger, 2012b). Demzufolge dürfte die Arbeitsunzufriedenheit auch im Jahr 2011 auf dem Niveau der Vorjahre stagnieren (Hypothese 4). Entsprechend den Analysen von Appius (2012) ist zudem davon auszugehen, dass das Erleben von Unsicherheit, Entlastung, Leistungsdruck und Arbeitsunzufriedenheit in einer systematischen Beziehung zueinander steht, wobei Unsicherheit, Leistungsdruck und Arbeitsunzufriedenheit positiv korreliert sind, das Erleben von Entlastung durch das Zentralabitur jedoch in einer negativen Relation zu den anderen Indikatoren steht. Die Entlastung ist demnach umso höher, je geringer Unsicherheit, Leistungsdruck und Arbeitsunzufriedenheit ausfallen (Hypothese 5).

Neben der individuellen und kollektiven Erfahrung mit dem Zentralabitur sollten weitere Faktoren bezüglich des emotionalen Erlebens eine Rolle spielen. Der Befund von Oerke, dass „die Kooperation im Abitur und die kollektive Selbstwirksamkeit für die Auseinandersetzung mit der Reform von Bedeutung sind, spricht dafür, dass viele Probleme, die durch das Zentralabitur entstehen, in Zusammenarbeit mit anderen Lehrpersonen besser gelöst werden können als allein“ (2012b, S. 231). Dementsprechend und in Anlehnung an die Ergebnisse u.a. von Appius (2012), Bieri (2006), Fussangel und Gräsel (2011), Fussangel et al. (2010), Gehrmann (2007) sowie Soltau und Mienert (2010) wird angenommen, dass die abiturbezogene Kooperation, die kollektive Selbstwirksamkeit und das Schulklima zu einer Reduzierung der Unsicherheit, des Leistungsdrucks und der Arbeitsunzufriedenheit beitragen sowie die wahrgenommene Entlastung fördern (Hypothese 6).

5. STUDIE „IMPLEMENTATION UND AUSWIRKUNGEN NEUER STEUERUNGSSTRUKTUREN IM SCHULWESEN AM BEISPIEL ZENTRALER ABITURPRÜFUNGEN“

Die dem Beitrag zugrundeliegende Studie fand an der Universität Zürich in Kooperation mit dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF) statt. Sie wurde in den Jahren 2007 bis 2009 sowie im Jahr 2011 durchgeführt. In Bremen erfolgte das Zentralabitur 2011 zum fünften Mal, was die Untersuchung längerfristiger Effekte der Einführung erlaubt.

Design

Ziel der Studie ist die Analyse der „Implementation zentral organisierter Abiturprüfungen als ein Element im neuen Konzept der Systemsteuerung in den zwei deutschen Bundesländern Bremen und Hessen […]. Im Zentrum stehen Fragen a) zu den Effekten des Wechsels von einem dezentralen zu einem zentralen Prüfungssystem in Bremen für Schüler/-innen, Lehrpersonen, Unterricht und Schule sowie b) zu den Veränderungen des schulischen Handelns und der schulischen Leistungen nach Implementation zentraler Abiturprüfungen in beiden Bundesländern" (Maag Merki, 2012, S. 13). Hierfür wurden in den Jahren 2007, 2008, 2009 und 2011 Erhebungen bei Schülerinnen und Schülern, Lehrpersonen und Schulleitungen (nur 2011) durchgeführt. Einen Überblick über das Design und die eingesetzten Instrumente bietet Tabelle 1.

Tabelle 1: Forschungsdesign

	2007 bis 2009	**2011**
Schulen (N)	37 (Bremen: 19, Hessen: 18)	28 (Bremen: 19, Hessen: 9)
Schülerinnen und Schüler	• standardisierte Befragung vor und nach dem Abitur • Leistungstests in Mathematik und Englisch • KFT 4-12+R • Abiturpunktzahl/ Halbjahresnoten	• standardisierte Befragung vor dem Abitur • Leistungstest in Mathematik • KFT 4-12+R • Abiturpunktzahl/ Halbjahresnoten
Lehrpersonen	standardisierte Befragung vor und nach dem Abitur	standardisierte Befragung vor dem Abitur
Schulleitungen		standardisierte Befragung nach dem Abitur

Methodik

In diesem Beitrag werden nur die Entwicklungen in Bremen analysiert. Um die Übersichtlichkeit der Ergebnisse zu wahren, gehen in die folgenden Berechnungen lediglich die Daten der Jahre 2007, 2009 und 2011 ein. Detaillierte Auswertungen der Jahre 2007, 2008 und 2009 bietet Maag Merki (2012). Die hier interessierenden Facetten des emotionalen Erlebens des Zentralabiturs bei Bremer Lehrkräften wurden in allen Befragungsjahren mittels eines standardisierten Fragebogens jeweils vor dem schriftlichen Abitur erhoben. Die Skalierung der Items reichte von 1 = trifft gar nicht zu bis 4 = trifft genau zu.

- Unsicherheit gegenüber dem Zentralabitur (Skala aus vier Items: Cronbachs Alpha: 2007: $\alpha = .71$, 2009: $\alpha = .71$, 2011: $\alpha = .69$)
  Beispielitem: *Ich habe Angst, dass ein Thema kommt, in dem die Schüler/innen nicht gut vorbereitet sind.*
- Leistungsdruck durch das Zentralabitur (Einzelitem)
  Seit das Zentralabitur eingeführt ist, fühle ich einen größeren Leistungsdruck.
- Entlastung durch das Zentralabitur (Einzelitem)
  Das Zentralabitur hat mich in meiner bisherigen Arbeit entlastet.
- Arbeitsunzufriedenheit (Skala aus sechs Items: Cronbachs Alpha: 2007: $\alpha = .83$, 2009: $\alpha = .82$, 2011: $\alpha = .82$)
  Beispielitem: *Ich habe mir schon ernsthaft überlegt, aus dem Beruf auszusteigen.*

Auf der Schulebene sind das Schulklima, die Kooperation unter den Lehrpersonen sowie deren kollektive Selbstwirksamkeit von Interesse.

- Schulklima (Skala aus neun Items: Cronbachs Alpha: 2007: $\alpha = .88$, 2009: $\alpha = .88$, 2011: $\alpha = .89$)
  - Beispielitem: *Die Stimmung an unserer Schule ist meistens heiter/fröhlich – gedrückt/lustlos.*
  - Antwortskalierung: Ein hoher Wert (maximal 5) drückt eine positive Einschätzung aus, ein niedriger (minimal 1) eine negative.
- Kooperation in Zusammenhang mit dem Abitur (Skala aus sieben Items: Cronbachs Alpha: 2007: $\alpha = .81$, 2009: $\alpha = .82$, 2011: $\alpha = .81$)
  - Beispielitem: *Wie häufig treten bei Ihnen relativ regelmäßig die folgenden Handlungen im Zusammenhang mit der Vorbereitung auf das Abitur auf? Gemeinsame Besprechung von Sorgen und Problemen, die wir in Bezug auf das Abitur haben.*
  - Antwortskalierung: 1 = gar nicht, 2 = einmal im Jahr, 3 = mehrmals im Halbjahr, 4 = einmal monatlich, 5 = einmal wöchentlich
- Kollektive Selbstwirksamkeit (Skala aus fünf Items: Cronbachs Alpha: 2007: $\alpha = .70$, 2009: $\alpha = .73$, 2011: $\alpha = .76$)
  - Beispielitem: *Auch mit außergewöhnlichen Vorfällen können wir zurechtkommen, da wir uns im Kollegium gegenseitig Rückhalt bieten.*
  - Antwortskalierung: 1 = trifft gar nicht zu, 4 = trifft genau zu

Auswertungsstrategien

Zur Beantwortung der Frage, wie Bremer Lehrpersonen längerfristig emotional auf die Implementation des Zentralabiturs reagieren, werden zunächst mit den Querschnittsstichproben die vier Indikatoren Unsicherheit, Leistungsdruck, Entlastung und Arbeitsunzufriedenheit deskriptiv ausgewertet. Korrelationen dienen der Aufklärung des Zusammenhangs zwischen diesen vier Dimensionen. Deren Entwicklung über die Zeit wird einerseits im Querschnitt über die Berechnung von Effektstärken (Cohen, 1988) und andererseits im Längsschnitt mittels einfaktorieller Varianzanalysen mit Messwiederholung nachgezeichnet. Regressionsanalysen sollen die Auswirkungen von Erfahrung mit dem Zentralabitur (in Jahren), demographischem Hintergrund (Lehrerfahrung über Dienstalter[2], Geschlecht mit der Codierung 0 = männlich, 1 = weiblich) sowie von schulischen Merkmalen wie Schulklima, kollektiver Selbstwirksamkeit und Kooperation bezüglich des Zentralabiturs auf jedes dieser vier Konstrukte im Jahr 2011 ermitteln. Aufgrund der geringen Stichprobengröße im Längsschnitt (siehe Kapitel 5.3) wird hierfür auf die Querschnittsdaten zurückgegriffen.

Stichprobe

In Bremen fand eine schrittweise Implementation des Zentralabiturs statt: Im Jahr 2007 wurden zunächst die Grundkurse zentral, die Leistungskurse weiterhin dezentral geprüft. Im Jahr 2008 erfolgte die Umstellung ebenfalls in den Leistungskursen der Fächer Deutsch, Mathematik, Naturwissenschaften und fortgesetzte Fremdsprache, alle anderen Leistungskurse werden nach wie vor dezentral geprüft. Die unterschiedlichen Stichprobengrößen (vgl. Tabelle 2) lassen sich wie folgt erklären: Während alle Lehrpersonen die Fragen zu Arbeitsunzufriedenheit, Schulklima und kollektiver Selbstwirksamkeit gestellt bekamen, wurden die zur Kooperation bezüglich des Zentralabiturs nur von jenen beantwortet, die im jeweiligen Jahr eine 12. oder 13. Klasse unterrichteten. Im Gegensatz dazu waren die übrigen Fragen (Unsicherheit, Leistungsdruck und Entlastung) nur für diejenigen Lehrpersonen konzipiert, die beim (Zentral-)Abitur involviert waren.

[2] Da Dienstalter und Lebensalter sehr hoch korrelieren (r = .82***), wird das Lebensalter durch das Dienstalter ausreichend repräsentiert und nicht zusätzlich in die Berechnungen einbezogen. Das Dienstalter ist folgendermaßen gruppiert: 1 = Referendar/in, 2 = 5 Jahre oder weniger, 3 = 6-10 Jahre, 4 = 11-20 Jahre, 5 = 21-30 Jahre, 6 = mehr als 30 Jahre.

Tabelle 2: Stichprobe der Lehrpersonen in Bremen nach Jahren (Kohorten und Längsschnitt)

	Kohorten			Längsschnitt
	2007	**2009**	**2011**	**2007 bis 2011**
Unsicherheit durch das Zentralabitur	403	314	292	62
Leistungsdruck durch das Zentralabitur	401	306	277	60
Entlastung durch das Zentralabitur	401	313	282	60
Arbeitsunzufriedenheit	611	422	420	85
Schulklima	614	424	427	
Kooperation bezüglich Zentralabitur	517	388	357	
Kollektive Selbstwirksamkeit	610	420	417	
Geschlecht	582	337	356	
Lehrerfahrung	580	332	357	

N aller befragten Lehrpersonen in Bremen: 2007: N = 686, 2009: N = 482, 2011: N = 427

Die Erfahrung mit dem Zentralabitur wurde für jedes Jahr über die Häufigkeit der Teilnahme daran ermittelt. Im Jahr 2007 bestand lediglich die Möglichkeit der erstmaligen Beteiligung oder nicht, 2009 konnten Lehrpersonen null, ein, zwei oder drei Jahre Erfahrung haben. 2011 reicht die Spanne schließlich von null bis fünf Jahren Erfahrung.

6. ERGEBNISSE

In einem ersten Schritt findet eine Analyse der Querschnittsdaten aus dem Jahr 2011 statt. Hierfür werden jeweils zunächst die einzelnen Facetten des emotionalen Erlebens des Zentralabiturs bei Bremer Lehrpersonen deskriptiv vorgestellt und ins Verhältnis zu den Vorjahren (2007 und 2009) gesetzt. Anschließend wird mittels Korrelationen der Zusammenhang zwischen den verschiedenen Indikatoren ermittelt.

Unsicherheit gegenüber dem Zentralabitur

Im Jahr 2011 verneinen die Bremer Lehrkräfte im Durchschnitt eher, dass sie sich gegenüber dem Zentralabitur unsicher fühlen (M = 2.02, SD = .55). Dieses Ergebnis deckt sich mit den Befunden aus den Vorjahren, die von einer kontinuierlichen Abnahme seit dem ersten Jahr der Implementation 2007 (M = 2.25, SD = .61) bis zum Jahr 2009 (M = 2.01, SD = .56) zeugen. Von 2009 zu 2011 stagniert der Wert auf einem niedrigen Niveau, sodass die Differenz von 2007 zu 2009 sowie von 2007 zu 2011 jeweils d = -.40***[3] beträgt. Das bedeutet, dass im Kohortenvergleich die Unsicherheit in den Jahren 2009 und 2011 deutlich geringer ausfällt als zu Beginn der Implementation des Zentralabiturs im Jahr 2007.

Die Berechnung einer einfaktoriellen Varianzanalyse mit Messwiederholung mit den Daten derjenigen Lehrpersonen, die an allen drei Erhebungen der Jahre 2007, 2009 und 2011 beteiligt waren, bestätigt den im Querschnitt nachgewiesenen Effekt der Zeit bei der Unsicherheit gegenüber dem Zentralabitur auch für den Längsschnitt (N = 62; Wilks' Lambda = .75, $F(2, 60) = 9.85$, $p < .001$, multivariates partielles $Eta^2 = .25$). Die empfundene Unsicherheit der Lehrkräfte verändert sich also über die Jahre signifikant und zwar sowohl von 2007 zu 2009 ($p < .001$) als auch von 2007 bis 2011 ($p < .05$), jedoch nicht von 2009 zu 2011.

Leistungsdruck durch das Zentralabitur

Bezüglich des Zentralabiturs empfinden die Lehrpersonen im Jahr 2011 eher keinen Leistungsdruck (M = 2.21, SD = .83). Während der ersten drei Jahre seit der Implementierung blieb der empfundene Leistungsdruck im Bereich des arithmetischen Mittels relativ konstant (2007: M = 2.49, SD = .88; 2009: M = 2.41, SD = .84; 2007–2009: d = -.09), erst im fünften Jahr ist eine Abnahme zu erkennen. Damit ist unter Berücksichtigung der Querschnittstichprobe eine Verringerung bis 2011 festzustellen (2009–2011: d = -.24**; 2007–2011: d = -0.33***).

Die einfaktorielle Varianzanalyse mit Messwiederholung für den Leistungsdruck ergibt keinen signifikanten Effekt der Zeit (N = 60; Wilks' Lambda = .91, $F(2, 58) = 2.88$, n.s., multivariates partielles $Eta^2 = .09$).

[3] Signifikanz: $^+$: $p < .10$; *: $p < .05$; **: $p < .01$; ***: $p < .001$

Damit unterscheiden sich die Resultate des Quer- und Längsschnitts, wobei bei letzteren der geringere Stichprobenumfang zu bedenken ist.

Entlastung durch das Zentralabitur

Im fünften Jahr seit der Einführung des Zentralabiturs nehmen die befragten Lehrkräfte in Bremen dieses eher als Entlastung wahr (M = 2.59, SD = .75). Diese Haltung liegt jedoch nur knapp über dem arithmetischen Mittel von 2.5, sodass weiterhin von einer großen Gruppe von Lehrpersonen auszugehen ist, die sich durch die zentralen Prüfungen (noch) nicht entlastet fühlen. Über die Jahre hinweg ist eine deutliche Zunahme der Entlastung zu verzeichnen (2007: M = 2.15, SD = .84; 2009: M = 2.51, SD = .83), was die Effektstärken von d = .43*** (2007–2009), d = .10 (2009–2011) und d = .55*** (2007–2011) belegen.

Der im Querschnitt festgestellte Befund, dass die entlastende Komponente des Zentralabiturs mit der Zeit zunimmt, zeigt sich ebenfalls im Längsschnitt (N = 60; Wilks' Lambda = .70, $F(2, 58) = 12.39$, $p < .001$, multivariates partielles $Eta^2 = .30$). Analog zum Unsicherheitsempfinden ist der Effekt auf Veränderungen von 2007 zu 2009 und von 2007 bis 2011 zurückzuführen (je $p < .001$).

Arbeitsunzufriedenheit

Den Aussagen zur Arbeitsunzufriedenheit stehen die Lehrkräfte 2011 eher ablehnend gegenüber, d.h. sie treffen für sie eher nicht zu (M = 1.82, SD = .50). Damit unterscheidet sich diese Aussage zwar zu den Vorjahren (2009–2011: d = -.20**; 2007–2011: d = -.19**), doch war auch dort das Niveau nicht hoch (2007: M = 1.92, SD = .56; 2009: M = 1.92, SD = .55; 2007–2009: d = .01). Dennoch weisen die Ergebnisse darauf hin, dass die Arbeitsunzufriedenheit insgesamt im Zeitraum von 2007 bis 2011 abgenommen hat, sich also unter Berücksichtigung der Querschnittdaten eine eher positive Entwicklung zeigt. Die Lehrpersonen scheinen alles in allem mit ihrem Beruf nicht unzufrieden zu sein.

Die im Querschnitt zwar signifikante, aber leichte Verringerung der Arbeitsunzufriedenheit über die fünf untersuchten Jahre findet sich im Längsschnitt nicht (N = 85; Wilks' Lambda = .99, $F(2, 83) = 0.58$, n.s., multivariates partielles $Eta^2 = .01$). Zwischen den Jahren sind keine signifikanten Veränderungen festzustellen.

Zusammenhang zwischen den Komponenten des Emotionalen Erlebens des Zentralabiturs

Für die Stichprobe der Bremer Lehrpersonen, die 2011 an der Erhebung beteiligt waren, lässt sich gemäß Tabelle 3 festhalten, dass insbesondere die Unsicherheit und der empfundene Leistungsdruck in einem engen Zusammenhang stehen (r = .56***). Je höher die wahrgenommene Unsicherheit ausfällt, desto eher wird der Leistungsdruck bejaht – und umgekehrt. Darüber hinaus stehen Unsicherheits- und Entlastungsempfinden in Relation (r = -.31***). Je weniger Leistungsdruck durch das Zentralabitur entsteht, desto größer ist die entlastende Wirkung der neuen Prüfungsform (r = -.40***).

Tabelle 3: Korrelation der Indikatoren des Emotionalen Erlebens von Lehrpersonen in Bremen (2011)

	Unsicherheit	Leistungsdruck	Entlastung	Arbeits-unzufriedenheit
Unsicherheit		.559***	-.307***	.283***
Leistungsdruck			-.396***	.166**
Entlastung				-.210***

Signifikanz: $^{+}$: p < .10; *: p < .05; **: p < .01; ***: p < .001

Der Zusammenhang zwischen den drei genannten Faktoren und der Arbeitsunzufriedenheit fällt etwas schwächer aus. Am stärksten ist er noch zwischen ihr und der Unsicherheit (r = .28***). Der von Maslach, Schaufeli und Leiter (2001) berichtete Zusammenhang zwischen Burnout – einer extremen Form bzw. Folge von Belastung – und Arbeitsunzufriedenheit findet sich auch hier in der schwachen Korrelation von Leistungsdruck und Arbeitsunzufriedenheit (r = .17***).

Regressionsanalysen

In die Regressionsanalysen werden diejenigen Lehrkräfte einbezogen, die im Jahr 2011 an der Erhebung beteiligt waren. Aufgrund des Forschungsdesigns schwankt die Kohortengröße zwischen N = 230 und N = 294.

Tabelle 4: Regressionsanalysen zum emotionalen Erleben des Zentralabiturs 2011 (Kohortenvergleich)

Prädiktoren	Standardisierter Koeffizient Beta			
	Unsicherheit	Leistungs-druck	Entlastung	Arbeitsun-zufriedenheit
Erfahrung mit dem Zentralabitur	-.206**	-.200**	.197**	.017
Schulklima	-.151*	-.076	-.007	-.347***
Lehrerfahrung	-.136^{+}	-.083	.040	.068
Kooperation bezüglich Zentralabitur	-.105^{+}	-.074	-.022	-.054
Kollektive Selbstwirksamkeit	-.052	-.024	.103	-.239***
Geschlecht	.020	-.007	-.033	-.021
Insgesamt	R^2 = .133	R^2 = .073	R^2 = .059	R^2 = .285

N: Unsicherheit: N = 241; Leistungsdruck: N = 230; Entlastung: N = 232; Arbeitsunzufriedenheit: N = 294;
Signifikanz: $^{+}$: p < .10; *: p < .05; **: p < .01; ***: p < .001; standardisierte Regressionskoeffizienten

In Bezug auf die Prädiktoren, welche die Unsicherheit gegenüber dem Zentralabitur, den Leistungsdruck und die Entlastung durch dieses sowie die Arbeitsunzufriedenheit beeinflussen, lassen sich sowohl Gemeinsamkeiten als auch Unterschiede feststellen.

Als Gemeinsamkeit ist festzuhalten, dass bei allen vier Dimensionen des emotionalen Erlebens das Geschlecht der Lehrpersonen im Jahr 2011 keine Rolle spielt. Die Kennwerte fallen jeweils sehr niedrig und nicht signifikant aus. Unterschiedliches Empfinden des Zentralabiturs lässt sich also nicht

auf Differenzen zwischen Lehrerinnen und Lehrern zurückführen. Eine weitere Gemeinsamkeit stellen mit Ausnahme der Unsicherheit gegenüber dem Zentralabitur der Einfluss der Kooperation sowie die Lehrerfahrung dar. Sowohl die Zusammenarbeit unter den Lehrpersonen als auch die individuelle Lehrerfahrung leisten keinen signifikanten Beitrag zur Reduzierung von Leistungsdruck und Arbeitsunzufriedenheit sowie zur Erhöhung der Entlastung durch das Zentralabitur. Bezüglich der Unsicherheit sind diese Wirkungszusammenhänge tendenziell signifikant, was bedeutet, dass Kooperation im Kollegium im Zusammenhang mit dem Abitur die Unsicherheit der einzelnen Lehrkräfte tendenziell reduziert bzw. dass bei Lehrpersonen, die weniger kooperieren, die Unsicherheit eher größer ausfällt ($\beta = -.105^{+}$). Darüber hinaus verringert eine langjährige Diensttätigkeit tendenziell die Unsicherheit durch die neue Prüfungsform bzw. sind Lehrkräfte mit weniger Dienstjahren eher stärker verunsichert ($\beta = -.136^{+}$).

Unterschiede sind vorrangig zwischen der Arbeitsunzufriedenheit und den anderen Dimensionen erkennbar. Tendenziell sind für erstere andere Faktoren von Bedeutung als für die Unsicherheit, den Leistungsdruck und die Entlastung. So beeinflusst die kollektive Selbstwirksamkeit diese nicht, während sie höchst signifikant die Arbeitsunzufriedenheit reduziert. Im Gegenzug spielt bei ihr der Umfang der individuellen Erfahrung der Lehrpersonen mit dem Zentralabitur keine Rolle, bei den übrigen jedoch eine hoch signifikante. Beim Leistungsdruck ($\beta = -.200^{**}$) und bei der Entlastung ($\beta = .197^{**}$) ist die Erfahrung mit dem Zentralabitur jeweils der einzige aussagekräftige Prädiktor. Lehrpersonen, die im Jahr 2011 aufgrund mehrmaliger Beteiligung beim Zentralabitur auf Erfahrung im Umgang mit diesem zurückgreifen können, empfinden durch dieses weniger Leistungsdruck und erleben es eher als Entlastung als Lehrkräfte, auf die das (eher) nicht zutrifft. Bei der Unsicherheit wirkt sich neben der Erfahrung ($\beta = -.206^{*}$) noch das Schulklima ($\beta = -.151^{*}$) signifikant aus, d.h. in Schulen mit einer guten Stimmung unter den Lehrpersonen sowie zwischen ihnen und den Schülerinnen und Schülern fühlen sich Lehrkräfte durch das Zentralabitur im Jahr 2011 weniger verunsichert. Für eine niedrige Arbeitsunzufriedenheit sind ein positives Schulklima ($\beta = -.347^{***}$) sowie eine hohe kollektive Selbstwirksamkeit entscheidend ($\beta = -.239^{***}$). Das bedeutet im Umkehrschluss, dass an Schulen, an denen ein eher negatives Schulklima herrscht, oder die Lehrpersonen das Kollegium als nur wenig selbstwirksam wahrnehmen, die Arbeitsunzufriedenheit höher ausfällt.

Obwohl die Prädiktoren im Jahr 2011 bei der Arbeitsunzufriedenheit einen guten Beitrag zur Varianzaufklärung (29 Prozent) leisten, bleibt dennoch ein großer Teil ungeklärt. Bei den anderen Dimensionen des emotionalen Erlebens vermögen die berücksichtigten Faktoren Erfahrung mit dem Zentralabitur, Schulklima, individuelle Lehrerfahrung, Kooperation in Bezug auf das Zentralabitur, kollektive Selbstwirksamkeit und Geschlecht der Lehrpersonen noch weniger Varianz zu erklären. Hier besteht weiterer Forschungsbedarf. Untersucht werden sollten weitere Einflussfaktoren auf Leistungsdruck und Entlastung, z.B. das unterrichtete Fach, die Schwerpunktthemen des Abiturs, die unterrichteten Kurse (Grundkurs versus Leistungskurs, Anzahl der Kurse im Zentralabitur), der Umfang der Unterrichtstätigkeit pro Woche, die Arbeitsbedingungen oder schulspezifische Merkmale. Auf die Unsicherheit könnten sich beispielsweise die individuelle Selbstwirksamkeit sowie ebenfalls die Schwerpunktthemen des Zentralabiturs auswirken.

7. DISKUSSION

Das emotionale Erleben des Zentralabiturs bei Bremer Lehrpersonen und dessen Entwicklung vom ersten Jahr der schrittweisen Implementation der neuen Prüfungsform bis fünf Jahre danach wird in der vorgestellten Studie mittels verschiedener Facetten beleuchtet. Dabei liegt in diesem Beitrag der Fokus einerseits auf Unsicherheit, Leistungsdruck und Entlastung aufgrund der zentralen Abiturprüfungen, andererseits auf der Arbeitsunzufriedenheit. In allen vier Dimensionen haben über die Zeit Entwicklungen stattgefunden, wenn auch unterschiedlicher Art.

Für die Unsicherheit ist bei Betrachtung sowohl des Quer- als auch des Längsschnitts eine Abnahme über die fünf Jahre von 2007 bis 2011 zu konstatieren, womit Hypothese 1 zu bestätigen ist. Diese fand jedoch insbesondere von 2007 zu 2009 statt (vgl. dazu Oerke, 2012a). Da bereits im ersten Jahr des Zentralabiturs der Wert der Unsicherheit unterhalb des arithmetischen Mittels der Skala liegt, kann nicht von einer generellen Verunsicherung der Lehrpersonen gegenüber der neu eingeführten Prüfungsform gesprochen werden (siehe auch Maag Merki, 2008). Dies könnte darin begründet sein, dass zunächst ausschließlich die Grundkurse zentral, die Leistungskurse jedoch weiterhin dezentral geprüft wurden. Die schrittweise Implementation bot den Lehrpersonen die Gelegenheit, sich zuerst

in Kursen mit weniger Gewicht bei der Gesamtabiturnote mit den zentralen Prüfungen auseinanderzusetzen. Diese Vermutung wird durch Oerkes (2012a) Befund für die Daten von 2007 bis 2009, dass in diesem Zeitraum die Unsicherheit und der Leistungsdruck in Hessen, wo das Zentralabitur im Jahr 2007 zeitgleich für die Grund- und Leistungskurse eingeführt wurde, höher und die Entlastung geringer als in Bremen ausfällt, gestützt. Dennoch müssten laut der Autorin weitere Erklärungsansätze in Betracht gezogen werden. Das gilt hier ebenfalls.

Die die Unsicherheit im Jahr 2011 bedingenden Faktoren spielen eine unterschiedliche Rolle. Den größten Einfluss hat die quantitative Erfahrung mit dem Zentralabitur: Je umfangreicher diese ist, desto weniger fühlen sich die Lehrpersonen verunsichert. Neben der direkten Erfahrung mit zentralen Prüfungen ist auch die Lehrerfahrung von Bedeutung. Da im Schuldienst unerfahrenere Lehrkräfte über weniger Handlungssicherheit verfügen (Oelkers, 2000; siehe auch Floden & Buchmann, 1993), geht ihre geringere Lehrerfahrung tendenziell mit einer höheren Unsicherheit einher. Dienstältere Lehrerinnen und Lehrer scheinen sich hingegen durch das Zentralabitur nicht so schnell aus der Ruhe bringen zu lassen und könnten mehr von ihrem Erfahrungswissen, auch im Umgang mit Reformen, profitieren. Dass dies nicht erst fünf Jahre nach der Implementation der Fall ist, zeigen zudem die Resultate von Oerke (2012a) für die ersten drei Jahre der Einführung zentraler Abschlussprüfungen. Laut Munthe (2001, S. 169) stehen Unsicherheit und routiniertes Verhalten in einem Zusammenhang: „Certainty has been found to be related to more non-routine behaviour, or flexible behaviour, and uncertainty to routine behaviour". Demnach hätten sich bei aller Erfahrung diejenigen Lehrkräfte, die bereits länger unterrichten, ihre Flexibilität bewahrt, was ihnen bei Veränderungen hilft. Mit Bezug zu Lange und Burroughs-Lange (1994) ist jedoch zu bedenken, dass Lehrerfahrung nicht einzig an der Anzahl der Dienstjahre abzulesen ist, sondern darüber hinaus noch andere Facetten zu berücksichtigen sind. „Using professional experiences as a source of understanding to enhance growth involves not just accumulation over time but qualitative change in teacher thinking and actions" (ebd., S. 624).

Neben der Erfahrung mit dem Zentralabitur und dem schulischen Leben im Allgemeinen reduziert ein positives Schulklima die erlebte Unsicherheit. Eine freundliche, vertrauensvolle Atmosphäre vermittelt demnach in von Reformen geprägten Situationen Sicherheit. Einen tendenziell signifikanten Einfluss hat darüber hinaus die abiturbezogene Kooperation. Dies bestätigt in Teilen Hypothese 6 und deckt sich

mit den Ergebnissen von Oerke (2012a), die den Einfluss der Kooperation zwischen den Lehrpersonen auf die Unsicherheit unter Berücksichtigung der Erfahrung der Lehrpersonen und der Mehrebenenstruktur mit den längsschnittlichen Daten von 2007 bis 2009 untersucht. Appius (2012) stellt hingegen bei alleiniger Betrachtung des Jahres 2009 in einem Strukturgleichungsmodell keine Auswirkungen der Kooperation auf die Unsicherheit fest. Die Differenzen in den Ergebnissen könnten auf die verschiedene thematische Schwerpunktsetzung und die damit einhergehenden unterschiedlichen Analysemethoden, Kontrollvariablen und Stichproben zurückgehen. Die Autorinnen verweisen zudem darauf, dass weitere Analysen notwendig sind, um die Funktionalität von Kooperation zwischen Lehrpersonen auf das emotionale Erleben in langfristiger Perspektive differenziert beschreiben zu können. Die kollektive Selbstwirksamkeit beeinflusst entgegen Hypothese 6 die wahrgenommene Unsicherheit im Jahr 2011 nicht signifikant. Darüber hinaus sind keine geschlechtsspezifischen Differenzen auszumachen.

Die Unsicherheit steht entsprechend der formulierten Hypothese 5 in engem Zusammenhang mit dem empfundenen Leistungsdruck. Dieser nimmt zwar ebenfalls über die Zeit ab, jedoch erst nach 2009. Die Veränderung fällt lediglich im Querschnitt signifikant aus. Ob dies ausschließlich in der geringeren Stichprobengröße im Längsschnitt begründet ist, müssen weitere Analysen klären. Hypothese 2, wonach der Leistungsdruck über die Jahre abnimmt, ist somit für den Querschnitt zu bejahen, für den Längsschnitt allerdings abzulehnen. Im fünften Jahr der Einführung zentraler Abiturprüfungen mindert einzig eine mehrjährige Erfahrung mit diesen den Leistungsdruck. Die anderen in die Analysen einbezogenen Prädiktoren leisten entgegen Hypothese 6 keinen Beitrag zur Varianzaufklärung, sodass sie insgesamt lediglich bei sieben Prozent liegt. Bieri (2006) stellt geringe altersspezifische Differenzen bezüglich der Belastung zuungunsten der älteren Lehrpersonen sowie eine niedrigere Belastung der Lehrerinnen fest. Auch wenn der Einfluss der Lehrerfahrung hier nicht signifikant ist, weist er prinzipiell in die entgegengesetzte Richtung. Zudem sind keine Unterschiede zwischen Lehrerinnen und Lehrern erkennbar. Um die Merkmale, die mit einem geringen Belastungsempfinden einhergehen, zu identifizieren, bedarf es weiterer Forschung.

Jedes Jahr trägt das Zentralabitur etwas mehr zur Entlastung der Lehrpersonen bei, wodurch sich Hypothese 3 bestätigt. Dennoch liegt der Wert in 2011 lediglich knapp über dem arithmetischen Mittel, sodass im Durchschnitt eine Entlastung erst einige Jahre nach der Einführung festzustellen ist. Ähnlich

wie beim Leistungsdruck stellt die Erfahrung mit dem Zentralabitur die einzig signifikante Einflussgröße dar. Mit der Zeit gelangen die Lehrkräfte zu einem routinierteren Umgang mit den zentralen Prüfungen und gewinnen Handlungssicherheit. Dem Verlust der Kontrolle beispielsweise über die Abituraufgaben steht der Wegfall von deren zeit- und arbeitsintensiver Erstellung und damit von Verantwortung gegenüber. Dadurch verfügen die Lehrpersonen über mehr Ressourcen, um sich auf andere Facetten des (Schul-)Lebens zu konzentrieren. Ein Grund für das niedrige Niveau der Entlastung könnte die erhöhte Überprüfbarkeit der eigenen Arbeit der Lehrpersonen sein. Zwischen Leistungsdruck und Entlastung besteht erwartungskonform (Hypothese 5) ein negativer Zusammenhang: Je höher ersterer ausfällt, desto geringer wird die Entlastung bewertet.

Die Facetten des emotionalen Erlebens Unsicherheit, Leistungsdruck, Entlastung und Arbeitsunzufriedenheit korrelieren gemäß Hypothese 5 miteinander. Der Zusammenhang der Unsicherheit mit der Entlastung und Arbeitszufriedenheit sowie die Beziehung der beiden letztgenannten Indikatoren decken sich einerseits mit den Befunden für die Daten des Jahres 2009 (Appius, 2012). Andererseits steht die Relation von Arbeitsunzufriedenheit, Leistungsdruck und Entlastung in Übereinstimmung mit den Ergebnissen von Rudow (1994) zum Zusammenhang einer geringen Arbeitszufriedenheit mit Ermüdung und Stress. Dennoch ist die Arbeitsunzufriedenheit von Lehrpersonen von anderen Faktoren beeinflusst als die Unsicherheit, der Leistungsdruck und die Entlastung. Dies könnte darin begründet sein, dass die Arbeitsunzufriedenheit im Gegensatz zu den übrigen drei Indikatoren nicht direkt mit Bezug auf das Zentralabitur, sondern allgemein erhoben wurde. Der schulische Kontext – erfasst über das Schulklima und die kollektive Selbstwirksamkeit – spielt vorrangig eine Rolle, nicht aber die Erfahrung mit dem Zentralabitur. So reduzieren ein positiv empfundenes Schulklima sowie beispielsweise die Überzeugung, Herausforderungen gemeinsam bewältigen zu können, die Arbeitsunzufriedenheit, was teilweise Hypothese 6 belegt. Zudem besteht bezüglich des positiven Einflusses des Schulklimas Übereinstimmung zu den Befunden von Appius (2012) für das Jahr 2009. Die Verringerung der Arbeitsunzufriedenheit über die fünf Jahre ist im Querschnitt zwar von signifikanter Größe, im längsschnittlichen Vergleich jedoch nicht. Damit ist Hypothese 4 tendenziell zu bestätigen. Die Arbeitsunzufriedenheit scheint vielmehr zeitlich relativ stabil zu sein. Zu diesem Schluss kommen sowohl Jäger (2012b) mit den Daten für die Jahre 2007 bis 2009 als auch Gehrmann (2007, S. 199): „Insgesamt verweisen die Studien über die berufliche Zufriedenheit von Lehrern auf zeitlich sehr stabile Einstellungen und Erfahrungen, die es den Lehrkräften ermöglichen, dauerhaft ihre Tätigkeit zu bewältigen". Laut Gehrmann

(2007) unterscheidet sich die Arbeitsunzufriedenheit weder geschlechts- noch altersspezifisch. Das lässt sich hier aufgrund der jeweils nicht signifikanten Einflüsse ebenfalls festhalten und steht damit im Gegensatz zu den Analysen von Bieri (2006) und Jäger (2012b). Laut Jäger (2012b) sind in den Jahren 2007 bis 2009 einerseits sowohl Lehrpersonen mit weniger als sechs Jahren Schuldienst als auch Lehrerinnen zufriedener, andererseits geht mehr Erfahrung mit dem Zentralabitur tendenziell mit einer größeren Arbeitsunzufriedenheit einher. Potentielle Erklärungen sind im Zuge der Implementation des Zentralabiturs entstandene Belastungen, Einschränkungen der Handlungsspielräume sowie eine niedrigere Bewertung der gesellschaftlichen Position (ebd.). Diese Faktoren haben sich fünf Jahre nach der Einführung möglicherweise abgeschwächt, sodass kein signifikanter Einfluss der Erfahrung mit dem Zentralabitur mehr festzustellen ist.

Hypothese 6, wonach die Kooperation bezüglich des Abiturs, die kollektive Selbstwirksamkeit und das Schulklima die Unsicherheit, den Leistungsdruck und die Arbeitsunzufriedenheit reduzieren sowie die Entlastung fördern, kann aufgrund divergierender Ergebnisse lediglich teilweise angenommen werden. So beeinflusst ein gutes Schulklima zwar die Unsicherheit und die Arbeitsunzufriedenheit günstig, hat jedoch keine Auswirkungen auf den Leistungsdruck und das Entlastungsempfinden. Eine hohe kollektive Selbstwirksamkeit trägt ausschließlich zu einer geringen Arbeitsunzufriedenheit bei und die Kooperation vermag nur tendenziell die Unsicherheit zu verringern. Appius (2012) zeigt für das Jahr 2009, dass jüngere Lehrpersonen öfter kooperieren. Ob dies im Jahr 2011 ebenfalls zutrifft und möglicherweise das Dienstalter über die Kooperation moderiert wird und deshalb sein Einfluss – mit Ausnahme bei der Unsicherheit – nicht signifikant ausfällt, müssen weitere Analysen klären. Darüber hinaus ist zu bedenken, dass andere Aspekte der Kooperation, d.h. in Bezug auf Benotung, Unterricht und Curriculum, sowie ein positives Schulklima die abiturbezogene Kooperation beeinflussen (Appius, 2012), was hier nicht berücksichtigt werden konnte.

Zusammenfassend bleibt festzuhalten, dass sich das emotionale Erleben des Zentralabiturs bei Bremer Lehrpersonen im Zeitraum von fünf Jahren seit Beginn dieser Prüfungsform in eine gute Richtung entwickelt hat: Unsicherheit und Leistungsdruck haben sich verringert und die Entlastung hat zugenommen, was sich auch in einer niedrigeren Arbeitsunzufriedenheit widerspiegelt. Dennoch gibt es weiterhin eine Gruppe von Lehrkräften, auf die dies (noch) nicht zutrifft.

8. AUSBLICK

Der vorliegende Beitrag nimmt den emotionalen Umgang mit dem Zentralabitur von Bremer Lehrpersonen fünf Jahre nach der Implementation dieser neuen Prüfungsform in den Blick und zeichnet dessen quer- und längsschnittliche Entwicklung seit dem Beginn im Jahr 2007 nach. Als Forschungsdesiderate müssen unter anderem eine mehrebenenanalytische Auswertung der vier Facetten Unsicherheit, Leistungsdruck, Entlastung und Arbeitsunzufriedenheit sowie der Einbezug von unterrichts- und fachspezifischen Merkmalen offen bleiben. Aufschlussreich wären darüber hinaus der Vergleich mit dem emotionalen Umgang der Schülerinnen und Schüler sowie das Aufdecken von Differenzen zwischen Bremen und Hessen, die im unterschiedlichen Implementationsmodus begründet sind. Die Forderung von Krause et al. (2011) nach längsschnittlichem Forschungsdesign ist erfüllt und bietet somit eine fundierte Grundlage für bereits durchgeführte und weitere Analysen, die für künftige Reformen von Nutzen sein können. Dies gilt auch für den Befund, dass mit einer Ausnahme die individuelle Erfahrung mit dem Zentralabitur in Zusammenhang mit den hier berücksichtigten Dimensionen steht: Sie trägt zur Reduzierung bzw. Kompensation von Unsicherheit und Leistungsdruck bei und begünstigt die Wahrnehmung der entlastenden Wirkung des Zentralabiturs. Damit ist sie ein entscheidender Faktor für die Implementation und Integration zentraler Abiturprüfungen. Sie kann jedoch nicht „verordnet", sondern lediglich zu einem gewissen Grad durch Kooperation im Kollegium kompensiert werden. Oerke (2012a) zeigt mittels der Daten der Jahre 2007 bis 2009, dass die Bedeutung der Erfahrung mit dem Zentralabitur in Zusammenhang mit der dadurch verursachten Unsicherheit umso geringer ausfällt, je mehr Kooperation stattfindet. Dies sollte bei der Planung von Veränderungen im Bildungswesen, beispielsweise beim Gemeinsamen Kernabitur, bedacht werden. Dabei ist auf die Schaffung bzw. Erweiterung geeigneter Bedingungen für Kooperation im Kollegium zu achten. Eine Reform der Prüfungen betrifft nicht nur die Lehrpersonen als „Einzelkämpfer", sondern die ganze Schule: „Was eine Innovation für den eigenen Unterricht bedeutet und wie das eigene Handeln angepasst bzw. verändert werden muss, sind Fragen, die Lehrkräfte in kooperativen Prozessen erarbeiten können. Zudem ist die Realisierung einer Innovation an einer Schule keine individuelle Aufgabe, sondern erfordert eine Veränderung der gesamten Institution" (Fussangel & Gräsel, 2011, S. 675).

LITERATURVERZEICHNIS

Amrein, A. L. & Berliner, D. C. (2002). High-Stakes Testing, Uncertainty, and Student Learning. *Education Policy Analysis Archives, 10 (*18). Verfügbar unter http://epaa.asu.edu/ojs/article/view/297/423 [19.01.2012].

Appius, S. (2012). Kooperationen zwischen Lehrpersonen im Zusammenhang mit dem Abitur. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 91–113). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bieri, T. (2006). *Lehrpersonen: Hoch belastet und trotzdem zufrieden?* Bern et al.: Haupt.

Bishop, J. H. (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review, 6*, 349–398.

Block, R., Klein, E. D., van Ackeren, I. & Kühn, S. M. (2011). Leistungseffekte des Zentralabiturs? Eine kritische Auseinandersetzung mit bildungsökonomischen Interpretationen zu den Effekten der Prüfungsorganisationen auf der Basis von PISA-E-2003-Daten. *Bildungsforschung, 8* (1), 215–238.

Böhm-Kasper, O. (2004). *Schulische Beanspruchung und Belastung. Eine Untersuchung von Schülern und Lehrern am Gymnasium*. Münster et al.: Waxmann.

Böhm-Kasper, O. & Weishaupt, H. (2002). Belastung und Beanspruchung von Lehrern und Schülern am Gymnasium. *Zeitschrift für Erziehungswissenschaft, 5* (3), 472–499.

Brown, M., Ralph, S. & Brember, I. (2002). Change-linked work-related stress in British teachers. *Research in Education 67*, 1–13.

Buchwald, P. (2011). *Stress in der Schule und wie wir ihn bewältigen*. Paderborn: Ferdinand Schöningh.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New York: Erlbaum.

Ditton, H. (2007). Schulqualität – Modelle zwischen Konstruktion, empirischen Befunden und Implementierung. In J. van Buer & C. Wagner (Hrsg.), *Qualität von Schule. Ein kritisches Handbuch* (S. 83–92). Frankfurt am Main et al.: Peter Lang.

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Floden, R. E. & Buchmann, M. (1993). Between Routines and Anarchy: preparing teachers for uncertainty. *Oxford Review of Education, 19* (3), 373–382.

Fussangel, K. (2008). *Subjektive Theorien von Lehrkräften zur Kooperation. Eine Analyse der Zusammenarbeit von Lehrerinnen und Lehrern in Lerngemeinschaften.* Wuppertal. Verfügbar unter http://elpub.bib.uni-wuppertal.de/edocs/dokumente/fbg/paedagogik/diss2008/fussangel/ [20.01.2012]

Fussangel, K. & Gräsel, C. (2011). Forschung zur Kooperation im Lehrerberuf. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 667–682). Münster et al.: Waxmann.

Fussangel, K., Dizinger, V., Böhm-Kasper, O. & Gräsel, C. (2010). Kooperation, Belastung und Beanspruchung von Lehrkräften an Halb- und Ganztagsschulen. *Unterrichtswissenschaft, 38* (1), 51–67.

Gehrmann, A. (2007). Zufriedenheit trotz beruflicher Beanspruchungen? Anmerkungen zu den Befunden der Lehrerbelastungsforschung. In M. Rothland (Hrsg.), *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen* (S. 185–203). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hargreaves, A. (2005). Educational change takes ages: Life, career and generational factors in teachers' emotional responses to educational change. *Teacher and Teacher Education, 21* (8), 967–983.

Hargreaves, A. (2004). Inclusive and exclusive educational change: emotional responses of teachers and implications for leadership. *School Leadership & Management, 24* (2), 287–309.

Hargreaves, A. (1998). The emotional practice of teaching. *Teaching and Teacher Education, 14* (8), 835–854.

Helsper, W. (2000). Antinomien des Lehrerhandelns und die Bedeutung der Fallrekonstruktion – Überlegungen zu einer Professionalisierung im Rahmen universitärer Lehrerausbildung. In E. Cloer, D. Klika & H. Kunert (Hrsg.), *Welche Lehrer braucht das Land? Notwendige und mögliche Reformen der Lehrerbildung* (S. 142–177). Weilheim und München: Juventa.

Herzog, S. (2007). *Beanspruchung und Bewältigung im Lehrerberuf.* Münster et al.: Waxmann.

Jäger, D. J. (2012a). Herausforderung Zentralabitur: Unterrichtsinhalte variieren und an Prüfungsthemen anpassen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 175–201). Wiesbaden: VS Verlag für Sozialwissenschaften.

Jäger, D. J. (2012b). Schulklima, Selbstwirksamkeit und Arbeitszufriedenheit aus Sicht der Lehrpersonen und Schüler/-innen in Hessen und Bremen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 61–89). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kamski, I. (2011). *Innerschulische Kooperation in der Ganztagsschule. Eine Analyse der Zusammenarbeit von zwei Berufsgruppen am Beispiel von Lehrkräften und Erzieherinnen und Erziehern.* Münster et al.: Waxmann.

Kelchtermans, G. (2005). Teachers' emotions in educational reforms: Self-understanding, vulnerable commitment and micropolitical literacy. *Teaching and Teacher Education, 21,* 995–1006.

Klusmann, U., Kunter, M. & Trautwein, U. (2009). Die Entwicklung des Beanspruchungserlebens bei Lehrerinnen und Lehrern in Abhängigkeit beruflicher Verhaltensstile. *Psychologie in Erziehung und Unterricht, 56*, 200–212.

Krapp, A. & Hascher, T. (2011). Forschung zu Lehreremotionen. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 511–526). Münster et al.: Waxmann.

Krause, A., Dorsemagen, C. & Alexander, T. (2011). Belastung und Beanspruchung im Lehrerberuf – Arbeitsplatz- und bedingungsbezogene Forschung. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 788–813). Münster et al.: Waxmann.

Kühn, S. M. (2012). Zentrale Abiturprüfungen im nationalen und internationalen Vergleich mit besonderer Perspektive auf Bremen und Hessen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 25–42). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lange, J. D. & Burroughs-Lange, S. G. (1994). Professional uncertainty and professional growth: A case study of experienced teachers. *Teaching & Teacher Education, 10* (6), 617–631.

Lortie, D. C. (1975). *Schoolteacher – a sociological study*. Chicago: The University of Chicago Press.

Lüsebrink, I. (2002). Unsicherheit als Herausforderung. Ein Beitrag zur Professionalisierung des LehrerInnenberufs. *Die Deutsche Schule, 94* (1), 39–49.

Maag Merki, K. (Hrsg.). (2012). *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Maag Merki, K. (2008). Die Einführung des Zentralabiturs in Bremen – Eine Fallanalyse. *Die Deutsche Schule, 100*(3), 357–368.

Maslach, C., Schaufeli W. B. & Leiter, M. P. (2001). Job Burnout. *Annual Review of Psychology, 52*, 397–422. Verfügbar unter http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.52.1.397 [20.01.2012]

Munthe, E. (2003). Teachers' workplace and professional uncertainty. *Teaching and Teacher Education, 19*, 801–813.

Munthe, E. (2001). Measuring Teacher Certainty. *Scandinavian Journal of Educational Research, 45* (2), 167–181.

Nichols, S. L. & Berliner, D. C. (2007). *Collateral Damage. How High-Stakes Testing courrupts American's schools*. Cambridge: Harvard Education Press.

Oelkers, J. (2000). Probleme der Lehrerbildung: Welche Innovationen sind möglich? In E. Cloer, D. Klika & H. Kunert (Hrsg.), *Welche Lehrer braucht das Land? Notwendige und mögliche Reformen der Lehrerbildung* (S. 126–141). Weilheim und München: Juventa.

Oerke, B. (2012a). Emotionaler Umgang von Lehrkräften und Schüler/-innen mit dem Zentralabitur: Unsicherheit, Leistungsdruck und Leistungsattributionen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 115–149). Wiesbaden: VS Verlag für Sozialwissenschaften.

Oerke, B. (2012b). Auseinandersetzung der Lehrpersonen mit der Einführung des Zentralabiturs: Stages of Concern. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 203–232). Wiesbaden: VS Verlag für Sozialwissenschaften.

Pedulla, J., Abrams, L. M., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Chestnut Hill, MA: National Board on Educational Testing and Public Policy, Lynch School of Education, Boston College.

Reusser, K., Pauli, C. & Elmer, A. (2011). Berufsbezogene Überzeugungen von Lehrerinnen und Lehrern. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 478–495). Münster et al.: Waxmann.

Rothland, M. (Hrsg.). (2007a). *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Rothland, M. (2007b). Soziale Unterstützung. Bedeutung und Bedingungen im Berufsalltag von Lehrerinnen und Lehrern. In M. Rothland (Hrsg.), *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Rudow, B. (1990). Konzepte zur Belastungs- und Beanspruchungsanalyse im Lehrerberuf. *Zeitschrift für Pädagogische Psychologie, 4* (1), 1–12.

Rudow, B. (1994). *Die Arbeit des Lehrers. Zur Psychologie der Lehrertätigkeit, Lehrerbelastung und Lehrergesundheit*. Bern et al.: Hans Huber.

Ryan, K. E., Ryan, A. M., Arbuthnot, K., & Samuels, M. (2007). Students' Motivation for Standardized Math Exams. *Educational Researcher, 36* (1), 5-13.

Schaarschmidt, U. & Kieschke, U. (2007). Beanspruchungsmuster im Lehrerberuf. Ergebnisse und Schlussfolgerungen aus der Potsdamer Lehrerstudie. In M. Rothland (Hrsg.), *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen* (S. 81–98). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schwarzer, R. (2000). *Streß, Angst und Handlungsregulation* (4., überarbeitete Aufl.). Stuttgart: Kohlhammer.

Schwarzer, R. & Warner, M. (2011). Forschung zur Selbstwirksamkeit bei Lehrerinnen und Lehrern. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 496–510). Münster et al.: Waxmann.

Sieland, B. (2007). Wie gehen Lehrkräfte mit Belastungen um? Belastungsregulierung zwischen Entwicklungsbedarf und Änderungsresistenz. In M. Rothland (Hrsg.), *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen* (S. 206–226). Wiesbaden: VS Verlag für Sozialwissenschaften.

Soltau, A. & Mienert, M. (2010). Unsicherheit im Lehrerberuf als Ursache mangelnder Lehrerkooperation? Eine Systematisierung des aktuellen Forschungsstandes auf Basis des transaktionalen Stressmodells. *Zeitschrift für Pädagogik, 56* (5), 761–778.

Soltau, A. & Mienert, M. (2009). Teamorientierung und Einstellungen zu Formen der Lehrerkooperation bei Lehrkräften. *Psychologie in Erziehung und Unterricht, 56,* 213–223.

Stähling, R. (1998). *Beanspruchung im Lehrerberuf. Einzelfallstudie und Methodenerprobung.* Münster et al.: Waxmann.

Stöckli, G. (1992). *Reaktionen auf Unterrichtssituationen. Eine experimentelle Untersuchung zur Belastung von Lehrerinnen und Lehrern der Mittelstufe.* Zürich: Pädagogisches Institut der Universität Zürich. Verfügbar unter http://www.ife.uzh.ch/pp1/stoeckli/zur_Person_files/Reaktionen.pdf [19.01.2012].

van Dick, R. & Stegmann, S. (2007). Belastung, Beanspruchung und Stress im Lehrerberuf – Theorien und Modelle. In M. Rothland (Hrsg.), *Belastung und Beanspruchung im Lehrerberuf. Modelle, Befunde, Interventionen* (S. 34–51). Wiesbaden: VS Verlag für Sozialwissenschaften.

vbw – Vereinigung der Bayerischen Wirtschaft e.V. (Hrsg.). (2011). *Gemeinsames Kernabitur. Zur Sicherung von nationalen Bildungsstandards und fairem Hochschulzugang*. Gutachten. Münster et al.: Waxmann.

Publikation 4: Publikation Emotionales Erleben des Zentralabiturs von Schülerinnen und Schülern

Maué, E. (2017). Die Implementation zentraler Abiturprüfungen und deren potentielle Auswirkungen auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg von Schülerinnen und Schülern. Zeitschrift für Pädagogik, 63(6), 803-826.

Zusammenfassung

Emotionen beeinflussen im Zusammenspiel mit weiteren Aspekten das Lernen und Lehren, wobei die Interaktion von Person und Umwelt von großer Bedeutung ist. Vor diesem Hintergrund analysiert der Beitrag mittels eines Kohortenvergleichs über fünf Jahre mögliche kurz- und längerfristige Auswirkungen der Implementation des Zentralabiturs auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg von Abiturientinnen und Abiturienten sowie auf den Einfluss unterrichtlicher und schulischer Faktoren auf diese Emotionen, zusätzlich nach Leistungsniveau differenziert. Strukturgleichungsmodelle zeigen längerfristig (2007-2011) die Stabilität der Emotionen, eine Abnahme des Effekts des Schulklimas auf die Erfolgsunsicherheit im Abitur sowie differenzielle Wirkungen je nach Leistungsniveau. Die meisten Effekte bleiben jedoch über die Jahre und die Prüfungsform (dezentral-zentral) konstant.

Summary

Emotions, combined with other aspects, influence the interaction between a person and his/her environment, learning, and teaching. Comparisons of cohorts over five years examine potential short- and long-term effects not only on students' uncertainty of success and anxiety of failure caused by the implementation of state-wide exit exams at the end of upper secondary education, but also on the impact of teaching, differentiated by students' achievement level. In a long-term perspective (2007-2011), structural equation modeling shows a high stability of both emotions mentioned above, a decreasing influence of the school climate on the uncertainty of success, and differential effects of students' achievement level. However, regardless of the exam system (course-based vs. state-wide), most effects remain stable over the years.

1. EINLEITUNG

Emotionen spielen im Leben und damit auch im schulischen Kontext eine zentrale Rolle: „Emotions are dynamic parts of ourselves, and whether they are positive or negative, all organizations, including schools, are full of them" (Hargreaves, 1998, S. 835). Emotionen wie Angst oder Ängstlichkeit von Schülerinnen und Schülern, vor allem in Verbindung mit Leistung, scheinen gut erforscht (vgl. Goetz et al., 2004; Ma, 1999; Schnabel, 1998). Andere Emotionen wie Ärger, Langeweile oder Freude, das Zusammenspiel mit Motivation sowie weitere Akteure und Konstellationen, die das emotionale Erleben beeinflussen, stehen ebenfalls im Fokus der Forschung (vgl. Frenzel, Pekrun & Goetz, 2007; Hagenauer & Hascher, 2014; Pekrun & Linnenbrink-Garcia, 2014; Reyes, Brackett, Rivers, White & Salovey, 2012; Seifried, 2009). Offen ist hingegen weitgehend, ob sich im Mehrebenensystem Schule (Fend, 2008) durch Reformen auf Systemebene initiierte Veränderungen der Rahmenbedingungen von Schule und Unterricht auf das emotionale Befinden von Lernenden, auch unter Berücksichtigung ihres Leistungsniveaus, auswirken. Dies soll anhand eines Kohortenvergleichs im Zeitraum von 2007 bis 2011, in welchem die Umstellung von dezentralen auf zentrale Abiturprüfungen in Bremen stattfand, untersucht werden. Die Implementation erfolgte 2007 zunächst in allen Grundkursen, 2008 zusätzlich in den Leistungskursen Deutsch und Fremdsprachen sowie in Mathematik und den naturwissenschaftlichen Fächern. In den Leistungskursen der übrigen Fächer gibt es weiterhin dezentrale, von den Lehrpersonen erstellte Abiturprüfungen. Für die zentralen Prüfungen entwickelt eine externe Kommission die Aufgaben und Korrekturkriterien. Sowohl bei dezentralen wie zentralen Abiturprüfungen nimmt die Erstkorrektur die Kurslehrperson und die Zweitkorrektur eine Lehrperson derselben Schule vor. Die Ergebnisse der Abiturientinnen und Abiturienten haben vorrangig für deren (berufliches) Leben Konsequenzen *(high stakes)*, weniger für die Lehrpersonen und Schulen (z. B. keine Lohnwirksamkeit oder Schulschließungen wie im angloamerikanischen Raum; vgl. Fend, 2011, S. 18). Damit haben zentrale Abiturprüfungen in Deutschland im internationalen Vergleich insgesamt einen *low stakes*-Charakter (vgl. Klein, Kühn, van Ackeren & Block, 2009).

Mit zentralen Abschlussprüfungen werden von bildungspolitischer Seite Ziele wie Sicherung und Steigerung der Leistungen sowie Transparenz und Vergleichbarkeit von Anforderungen, Bewertungen und Abschlüssen verbunden (Bishop & Wößmann, 2004; Kühn, 2010; van Ackeren,

Klemm & Kühn, 2015). Damit einhergehend wird ihnen ein höherer „Wert" zugeschrieben als dezentralen Abschlussprüfungen (vgl. Wößmann, 2003). Somit gewinnt das erfolgreiche Bestehen nicht nur für Schülerinnen und Schüler, sondern auch für Lehrpersonen (u. a. größere Transparenz ihrer Leistungen, gesteigerte Bedeutung der Kongruenz zwischen den zentralen Prüfungen und dem ihnen vorgelagerten Unterricht; vgl. Klein, 2016; Oerke, Maag Merki, Holmeier & Jäger, 2011) und andere direkt oder indirekt an Unterricht und Schule Beteiligte an Bedeutung. Dies kann u. a. zu vermehrtem Stress und Druck führen und sich auf die Unterrichtsgestaltung – Stichwort *teaching to the test* – auswirken (vgl. Bishop, 1999; Koretz, 2008; van Ackeren et al., 2015).

Den letzten Punkt aufgreifend, geht dieser Beitrag der Frage nach, ob sich im Kohortenvergleich Veränderungen im emotionalen Erleben von Schülerinnen und Schülern am Ende der Sekundarstufe II zeigen, die sich mit der Implementation des Zentralabiturs in Verbindung bringen lassen. Zudem wird untersucht, ob unterrichtliche und schulische Faktoren diese Emotionen unter den Bedingungen des Zentralabiturs stärker beeinflussen als bei dezentraler Prüfungsorganisation und ob sich differenzielle Wirkungen in Abhängigkeit des Leistungsniveaus ausmachen lassen.

Einer theoretischen und empirischen Verortung des Beitrages (2.) folgt die Ableitung der Forschungsfragen und Hypothesen (3.). Das Methodenkapitel (4.) bereitet die Grundlage für die Analysen (5.), welche anschließend diskutiert werden (6.). Einige Schlussbemerkungen (7.) runden den Artikel ab.

2. THEORETISCHER UND EMPIRISCHER HINTERGRUND

„*Emotionen* sind mehrdimensionale Konstrukte, die aus affektiven, physiologischen, kognitiven, expressiven und motivationalen Komponenten bestehen" (Frenzel, Goetz & Pekrun, 2015, S. 202). Dabei ist zwischen der situationsspezifischen Einschätzung (state) und der eher generelleren Ausprägung (trait) zu unterscheiden (Pekrun, 2006).

Im schulischen Kontext sind gemäß der *Kontroll-Wert-Theorie* von Pekrun (2006) *leistungsbezogene Emotionen* wie beispielsweise Angst, Ärger, Freude, Hoffnung(-slosigkeit), Langeweile und Stolz relevant,

die sowohl domänen- als auch fachspezifisch organisiert sein können. Leistungsbezogene Emotionen lassen sich zum einen nach ihrem Fokus auf Aktivitäten *(activity emotions)* versus auf deren Ergebnissen *(outcome emotions)* unterscheiden. Letztere werden zusätzlich differenziert nach retrospektiven, aktuellen oder prospektiven Emotionen wie beispielsweise Hoffnung auf Erfolg oder Furcht vor Misserfolg (zwei unabhängige Tendenzen der *Leistungsmotivation*; vgl. Heckhausen, 1963). Zum anderen kommen Differenzierungen nach dem Wert (positiv/negativ) und dem Grad der Kontrolle bzw. der Aktivierung hinzu. Bezüglich des *Einflusses individueller Faktoren* auf Emotionen zeigen Ahmed, van der Werf, Minnaert und Kuyper (2010), dass *competence* und *value beliefs* bzw. *appraisals* negativ mit negativen Emotionen (Ärger, Angst/Ängstlichkeit, Langeweile) und positiv mit positiven Emotionen (Freude, Hoffnung, Stolz) korrelieren (auch Hagenauer & Hascher, 2014). Für die *Entstehung von Emotionen* sind *appraisals*, das heißt „kognitive Einschätzungen von Situationen, Tätigkeiten oder der eigenen Person" (Frenzel et al., 2015, S. 212), und nicht die Situationen oder Tätigkeiten selbst entscheidend.

Diese Bewertungsprozesse eines Individuums stehen jeweils in Bezug zur *Umwelt*, im vorliegenden Fall zur schulischen Umwelt, d. h. *Lehrpersonen*, *Klasse*, *Unterricht* und *Schule*. Dies spiegelt sich im *Adaptable Learning Model* (Boekaerts, 1992, S. 382-383) in der Verknüpfung von 1. Anforderungen der Aufgabe und deren physischem, sozialem und didaktischem Kontext, 2. für die Aufgabe relevanten Kompetenzen, 3. *traits*, Selbstkonzept, Angst, kurz- und längerfristigen Zielen, und von 4. *appraisals* wider. Neben den personenbezogenen Merkmalen, wie Kompetenzen, Motivation, Attributionen, Copingstrategien und *appraisals* (2.-4.), spielen die Lehrpersonen und ihr Unterricht, eingebettet in die Bedingungen der Klasse (z. B. Leistungsniveau; vgl. Goetz et al., 2004) und der Schule, als didaktischer und sozialer Kontext eine Rolle (1.). Demnach sind emotionale Reaktionen der Lehrpersonen oder Peers auf ein Leistungsergebnis (*interpersonale Theorie der Motivation*; vgl. Weiner, 2000) ebenso von Bedeutung wie die Lernbedingungen. Dieses Zusammenspiel wird ebenfalls in der Selbstbestimmungstheorie von Deci und Ryan (1993, S. 229) mit der Verbindung des Bedürfnisses nach Kompetenz oder Wirksamkeit (*effectance*), nach Autonomie oder Selbstbestimmung und nach sozialer Eingebundenheit (*social relatedness*) oder sozialer Zugehörigkeit (*affiliation*) betont. Zusätzlich ist die persönliche Relevanz bestimmter Handlungen wichtig, die im vorliegenden Fall mit dem erfolgreichen Absolvieren der den Schülerinnen und Schülern bevorstehenden Abiturprüfungen als Anforderungen der Aufgabe (1.) als hoch angesehen werden kann. Diese Relevanz fällt laut Wößmann (2003) bei zentralen Abschlussprüfungen als „Währung" des Bildungssystems höher aus als bei dezentralen Prüfungen.

Übersteigen die Anforderungen jedoch die persönlichen Ressourcen zur Bewältigung, besteht also ein Ungleichgewicht von „demands [concerning the student's intellectual, motivational, and social capability and capacity] stemming from the individual himself, from significant others, or from the school" (Rost & Schermer, 1987, S. 227), entsteht *Test-, Leistungs- oder Prüfungsangst*. Dabei wird zwischen den Komponenten *worry* (Besorgtheit; kognitive Aspekte) und *emotionality* (Aufgeregtheit; physiologische, affektive Aspekte) differenziert (Hodapp, 1991).

Hinsichtlich des *Zusammenspiels von Angst und (Schul-)Leistung* reichen die bisherigen Befunde von Leistungseinbußen aufgrund von Selbstbezug (Selbstzweifel, Antizipation von Misserfolg; vgl. Frenzel et al., 2015), über kein Zusammenhang (vgl. Zimmer & Hocevar, 1994) bis zu Leistungssteigerungen (vgl. Peters-Häderle, 2006) und Angst reduzierende Wirkung höherer Leistungen (vgl. Hembree, 1990; Schnabel, 1998). Insgesamt überwiegen jedoch Befunde zu negativen Effekten (vgl. Ma, 1999; Schumacher, 2016; Seipp, 1990; Zeidner & Schleyer, 1998). Die Einflüsse differieren zudem nach *state*- oder *trait*-Angst, die sich ihrerseits in *Abhängigkeit der Leistungsstärke* der Schülerinnen und Schüler unterscheiden (vgl. Gläser-Zikuda & Mayring, 2003; Roos et al., 2015). Leistungsstarke Lernende zeigen günstigere (Entwicklungen ihrer) Emotionen als leistungsschwache (vgl. Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004).

Test-, Leistungs- oder Prüfungsangst ist im Zusammenspiel von biologischen Ursachen, Sozialisation, schulischem Kontext, Lernerfahrungen, Testsituation, Testwahrnehmung, Copingstrategien und Anpassungsleistungen begründet (*transactional model of test anxiety*; vgl. Zeidner, 1998; auch Schumacher, 2016). Neben außerschulischen Prädiktoren, wie Leistungserwartungen der Eltern oder soziale Unterstützung, nimmt Schumacher (2016, S. 159) die Unterrichtsqualität während der Prüfungsvorbereitung, Schwierigkeit des Unterrichtsstoffs, Transparenz der Prüfungssituation, Konkurrenzorientierung der Schülerinnen und Schüler, den Leistungsdruck sowie das Bestrafungsverhalten der Lehrpersonen als schulische Prädiktoren in ihr Prozessmodell schulischer Prüfungsangst auf.

Weitere Faktoren, die Angst vor oder in Prüfungen beeinflussen, sind u. a. das Gefühl fehlender Kontrolle, schlechter Vorbereitung, mangelnder Arbeits- und Lerntechniken („Repertoire-Unsicherheit"; vgl. Rost & Schermer, 1987), unklare oder hohe Leistungsanforderungen und -erwartungen, Unsicherheit bezüglich des Ergebnisses, (Miss-)Erfolgserwartungen, Leistungsdruck, das Leistungsniveau der Klasse

sowie die Motivation, eine gute Note zu bekommen (vgl. Frenzel et al., 2007; Goetz et al., 2004; Ryan, Ryan, Arbuthnot & Samuels, 2007; Schumacher, 2016; Seipp, 1990; Zeidner & Schleyer, 1998).

Dass nicht nur individuelle Merkmale und Faktoren der schulischen Umwelt, sondern auch Bedingungen der Systemebene entscheidend sind, verdeutlichen Befunde, dass *zentrale Prüfungen* oftmals, u. a. aufgrund der erhöhten Relevanz (vgl. Schumacher, 2016; Wößmann, 2003), mit größerer Angst, Unsicherheit und verstärktem Leistungsdruck einhergehen (vgl. Jürges, Schneider, Senkbeil & Carstensen, 2009). In Bremer Leistungskursen steigt die Erfolgsunsicherheit im Abitur von 2007 (dezentrale Abiturprüfungen) zu 2008 und 2009 (je zentrale Abiturprüfungen) an, wobei das Gefühl guter Vorbereitung im Unterricht diese reduziert (vgl. Oerke, 2012). Für das Fach Mathematik zeigen sich keine Veränderungen im Niveau der Erfolgsunsicherheit im Abitur und der Angst vor Misserfolg von 2007 bis 2009 (vgl. Maag Merki, 2012).

Lehrpersonen und ihre *Unterrichtsgestaltung* haben eine tragende Rolle für das Lernen der Schülerinnen und Schüler – Stichwort ‚Auf den Lehrer kommt es an' (vgl. Lipowsky, 2005; Zierer, 2015) – und sie können negative Emotionen sowohl verstärken als auch reduzieren (vgl. Hospel & Galand, 2016; Ledergerber, 2015; Maier, 2002; Oerke, 2012). Transparente Leistungsstandards, Leistungserwartungen und Bewertungskriterien, Rückmeldungen, eine gute Vorbereitung, Motivierungsfähigkeiten sowie ein positives Klassen- und Schulklima können negative Emotionen mildern (vgl. Frenzel et al., 2007; Gläser-Zikuda & Fuß, 2008; Hospel & Galand, 2016; Klein, 2016; Ledergerber, 2015; Reyes et al., 2012; Rost & Schermer, 1987; Zeidner, 1998; Zeidner & Schleyer, 1998). Die Auswirkungen dieser Faktoren differieren abhängig von der Unterrichtsgestaltung (vgl. Hugener, 2008; Seifried, 2009) sowie nach *Leistungsstand* der Lernenden (vgl. Muijs, Campbell & Kyriakides, 2005; Seifried, 2009; Vanlaar et al., 2016). So fällt z. B. das Erleben von Autonomie und von motivierender Unterstützung, aber auch von negativen Emotionen, im konstruktivistisch orientierten Unterricht stärker aus als im instruktionalen und teilweise im systemorientierten Mischtyp (vgl. Seifried, 2009).

Das reziproke Verhältnis zwischen den Emotionen aller Beteiligten wird in Ländern mit *zentralen Abschlussprüfungen* zusätzlich durch den für Lehrpersonen grösseren Druck und Stress (vgl. Bishop, 1999) beeinflusst. Dieser wirkt sich über den Unterricht (vgl. Klein, 2016; Maag Merki & Oerke, 2017)

und den Zusammenhang von Emotionen der Lehrpersonen mit denen der Schülerinnen und Schüler, deren Kognition und Motivation (vgl. Hargreaves, 1998; Jennings & Greenberg, 2009; Thiel, 2016) auf die Lernumgebung aus.

Dies verdeutlicht erneut die entscheidende Rolle, die Lehrpersonen, ihr Verhalten und ihre Gestaltung des Unterrichts bzw. dessen Bewertung für das emotionale Erleben der Lernenden spielen, wobei der vorliegende Beitrag mit der Untersuchung von Erfolgsunsicherheit im Abitur und Angst vor Misserfolg auf die Komponente *worry* (Besorgtheit) fokussiert.

3. FORSCHUNGSDESIDERAT, FRAGESTELLUNGEN UND HYPOTHESEN

Theoretisch und empirisch ist die Relevanz von Emotionen für Lernen und Lehren unbestritten. Bezüglich des Einflusses externer Abschlussprüfungen auf das emotionale Erleben sind die Befunde jedoch inkonsistent und die Wirkungen des vorgelagerten Unterrichts am Ende der Sekundarstufe II unzureichend berücksichtigt. Darüber hinaus fehlen Forschungsbefunde, welche die Umstellung von dezentralen zu zentralen Abiturprüfungen in kurz- und längerfristiger Perspektive berücksichtigen.

Dieses Desiderat nimmt der vorliegende Beitrag am Beispiel Bremen mittels folgender Fragestellungen in den Blick:

1. Wie gestaltet sich die Einschätzung von Erfolgsunsicherheit im Abitur und Angst vor Misserfolg von Abiturientinnen und Abiturienten über einen Zeitraum von fünf Jahren (2007-2011)? Bestehen Veränderungen, die sich kurz- und längerfristig mit der Implementation zentraler Abiturprüfungen in Verbindung bringen lassen?

Den Kurslehrpersonen sind zwar die Schwerpunktthemen der zentralen Abiturprüfungen bekannt, nicht jedoch die von einer externen Kommission erstellten Aufgaben. Insofern ist mit der Umstellung des Prüfsystems 2008 von dem Gefühl mangelnder Kontrolle, von nicht eindeutigen Leistungsanforderungen sowie einer Steigerung der „Repertoire-Unsicherheit" (Rost & Schermer, 1987) sowohl für

die Lehrpersonen als auch für die Lernenden auszugehen. Dies dürfte bei den Abiturientinnen und Abiturienten mit größerer Erfolgsunsicherheit im Abitur und Angst vor Misserfolg einhergehen. Diese Auswirkungen sollten v. a. kurzfristiger Natur (2007-2008) sein, da mit der Zeit kollektive Erfahrungen des schulischen Umfeldes im Umgang mit dem Zentralabitur sowie die zunehmende Vielfalt an verfügbaren Informationen und Materialien (z. B. Prüfungen vergangener Jahre) Unsicherheiten mildern.

2. Verändern sich die Effekte der von den Schülerinnen und Schülern eingeschätzten Unterrichtsgestaltung (Vorbereitung auf das Abitur im Unterricht, Autonomie- und Kompetenzunterstützung, Leistungserwartungen und Motivierungsfähigkeit der Lehrperson), des Schulklimas und der Halbjahresnote 13/1 auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg durch die Einführung zentraler Abiturprüfungen?

Lehrpersonen und ihre Unterrichtsgestaltung erhalten beim Zentralabitur aufgrund der Trennung von Unterricht und Prüfung und der damit erforderlichen Passung zwischen diesen für das emotionale Erleben von Lernenden ein größeres Gewicht (vgl. Klein, 2016; Oerke et al., 2011). Es wird angenommen, dass kurz- und längerfristig die Effekte genannter Faktoren auf die Emotionen unter den Bedingungen zentraler Abiturprüfungen stärker ausfallen als bei dezentralen.

3. Zeigen sich differenzielle Auswirkungen in Abhängigkeit des Leistungsniveaus der Abiturientinnen und Abiturienten?

Die Halbjahresnoten beinhalten verschiedene individuelle Leistungen und entsprechende Rückmeldungen der Lehrperson, die sich je nach Leistungsniveau unterschiedlich auswirken (vgl. Urhahne, 2015). Schülerinnen und Schüler mit schwächerer Leistung zeigen ungünstigere (Entwicklungen ihrer) Emotionen als leistungsstarke (vgl. Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004). Zudem gewinnt mit der Umstellung auf das Zentralabitur der den Prüfungen vorausgehende Unterricht größere Bedeutung für die erste Gruppe (vgl. Oerke et al., 2011). Dementsprechend werden Unterschiede im Niveau und in der Entwicklung der Emotionen sowie differenzielle Wirkungen der unterrichtlichen und schulischen Faktoren auf die Emotionen in Abhängigkeit der Leistungsgruppe vermutet (vgl. Muijs et al., 2005; Seifried, 2009; Vanlaar et al., 2016).

4. METHODIK

4.1 Datenbasis und Stichprobe

Die Daten entstammen einer Studie, welche die Implementation und die Auswirkungen zentraler Abiturprüfungen in Bremen und Hessen im Zeitraum 2007 bis 2009 sowie 2011 untersucht. Für den vorliegenden Beitrag ist Bremen besonders aussagekräftig, da aufgrund der dortigen schrittweisen Einführung des Zentralabiturs in den Leistungskursen der Fächer Deutsch, Fremdsprachen, Mathematik und Naturwissenschaften Daten von dezentralen (2007) und zentralen Abiturprüfungen (ab 2008) vorliegen. In Bremen wurde eine längsschnittliche Vollerhebung aller Schulen mit gymnasialer Oberstufe (n = 19) realisiert. Pro Schule füllten die Lernenden jeweils eines Grund- und Leistungskurses in Mathematik und Englisch standardisierte Fragebögen aus. Sie beantworteten die Fragen separat für ihre drei schriftlichen Prüfungsfächer (zwei Leistungskurse, ein Grundkurs; Dreifachstichprobe). Die Erhebungen fanden vor dem Abitur, nach Abschluss des Halbjahres 13/1 statt. Die Rücklaufquote betrug 51% in 2007 (n = 751), 65% in 2008 (n = 977) und 74% in 2011 (n = 1157). Da es sich jeweils um Abiturientinnen und Abiturienten handelt, ist auf Individualebene lediglich ein Kohortenvergleich möglich, nicht jedoch ein längsschnittliches Design.

In die Berechnungen gehen die Bremer Daten der Jahre 2007, 2008 und 2011 ein, um kurzfristige (2007-2008) und längerfristige (2007-2011) Auswirkungen der Implementation des Zentralabiturs zu berücksichtigen. Die Stichprobe umfasst auf Individualebene 4451 Lernende in Leistungskursen mit Wechsel von dezentralen zu zentralen Abiturprüfungen im Querschnitt: 2007 (dezentral): n = 1368, 2008 (zentral): n = 1497 und 2011 (zentral): n = 1586. Bei zwei Leistungskursen in Fächern mit Wechsel des Prüfsystems sind die Schülerinnen und Schüler in den Analysen doppelt enthalten (Maag Merki & Oerke, 2012, S. 52). Dies erklärt die im Vergleich zur Rücklaufquote größere Stichprobe. Fehlende Werte wurden mittels multipler Imputation ergänzt (Maag Merki & Oerke, 2012, S. 51); ausgenommen ist die Halbjahresnote 13/1, bei der die Stichprobe etwas geringer ausfällt: 2007: n = 1350; 2008: n = 1457; 2011: n = 1552; n_{gesamt} = 4359. Bezüglich der Verteilung der Fächerhäufigkeiten ist Repräsentativität gegeben (Maag Merki & Oerke, S. 60).

4.2 Variablen

Die Einschätzung unterrichtlicher und schulischer Faktoren seitens der Schülerinnen und Schüler ist für die Entstehung und Ausprägung von Emotionen, hier Erfolgsunsicherheit im Abitur und Angst vor Misserfolg, von Bedeutung. Die Auswahl der Skalen (Tab. 1) erfolgte vor dem theoretischen und empirischen Hintergrund des Zusammenspiels von Merkmalen der Person, der Lernumgebung sowie der sozialen Beziehung zwischen Lehrpersonen und Lernenden. Als Merkmal der Person wurde als Indikator für die individuelle Leistung (vgl. Boekaerts, 1992; Gläser-Zikuda & Mayring, 2003) die Note im Leistungskurs im Halbjahr 13/1 (Range: 0-15 Punkte) berücksichtigt. Die Bedeutung der Lernumgebung wurde mittels der Einschätzung der Schülerinnen und Schüler ihres Unterrichts hinsichtlich der Vorbereitung im Unterricht auf das Abitur (vgl. Oerke, 2012), der Autonomie- und Kompetenzunterstützung (vgl. Deci & Ryan, 1993; Hospel & Galand, 2016; Ledergerber, 2015) sowie der Motivierungsfähigkeit der Lehrpersonen (vgl. Gläser-Zikuda & Fuß, 2008; Seifried, 2009; Thiel, 2016) operationalisiert. Die wahrgenommenen Leistungserwartungen der Lehrperson spiegeln einen Aspekt der Leistungserwartungen der Umwelt (vgl. Goetz et al., 2004) wider. Das Schulklima steht für die soziale Beziehung zwischen Lehrpersonen und Lernenden (vgl. Jennings & Greenberg, 2009; Reyes et al., 2012).

Tabelle 1: In den Analysen verwendete Skalen

Skala	N Items	Beispielitem	Cronbach's α			Quelle
			2007	2008	2011	
Erfolgsunsicherheit im Abitur[a]	5	Ich bin besorgt, dass etwas im Abitur schief laufen könnte.	.83	.81	.81	Eigenentwicklung; Rakoczy, Buff & Lipowsky, 2005
Angst vor Misserfolg[a]	4	Wenn ich ein Problem nicht sofort verstehe, werde ich ängstlich.	.73	.74	.69	Grob & Maag Merki, 2001; Maag Merki, 2006
Vorbereitung im Unterricht[a]	3	Die möglichen Prüfungsthemen wurden im Unterricht ausführlich besprochen.	.80	.80	.82	Eigenentwicklung
Autonomie-unterstützung[a]	4	Im Unterricht habe ich die Möglichkeit, neue Themen selbstständig zu erkunden.	.68	.68	.65	Leutwyler & Maag Merki, 2005
Kompetenz-unterstützung[a]	4	Im Unterricht werde ich oft für gute Leistungen gelobt.	.81	.78	.76	Leutwyler & Maag Merki, 2005
Motivierungs-fähigkeit[a]	5	Im Unterricht steckt mich die Begeisterung meiner Lehrperson immer wieder an.	.82	.80	.82	Baumert, Gruehn, Heyn, Köller & Schnabel, 1997; Leutwyler & Maag Merki, 2005
Leistungs-erwartungen[a]	3	Unsere Lehrperson stellt hohe Anforderungen an uns.	.79	.75	.76	Maag Merki, 2002
Schulklima[b]	9	Die Stimmung an unserer Schule ist meistens… angstbesetzt – angstfrei.	.86	.84	.83	Eder, 1998; Leutwyler & Maag Merki, 2005

Anmerkungen: [a] Antwortmöglichkeiten: 1 = trifft gar nicht zu, 2 = trifft eher nicht zu, 3 = trifft eher zu bis 4 = trifft genau zu; [b] Antwortmöglichkeiten: 1 = negativ bis 5 = positiv

Zusätzlich werden die Analysen nach der Leistungsstärke, operationalisiert über die Note im Halbjahr 13/1, differenziert: Lernende mit starkem (oberes Quartil), mittlerem (mittlere Quartile) und geringem Leistungsniveau (unteres Quartil).

4.3 Analysen

Deskriptive Statistiken zeigen die über die zehn imputierten Datensätze gepoolten Mittelwerte, gepoolten Standardabweichungen und Standardfehler der Mittelwertschätzung (vgl. Rubin, 1987). Jahresvergleiche decken kurzfristige (2007-2008) und längerfristige (2007-2011) Auswirkungen auf. Deren Prüfung auf Signifikanz erfolgt mittels Regressionen der Variablen auf die Jahre sowie multiplen Gruppenvergleichen mit M*plus* Version 7.3 (vgl. Muthén & Muthén, 1998-2012). Effektstärken geben das Ausmass dieser Differenzen an (vgl. Cohen, 1988). Korrelationen belegen den Zusammenhang zwischen den Variablen Erfolgsunsicherheit im Abitur und Angst vor Misserfolg. Ob sich die Korrelationen signifikant zwischen den Jahren unterscheiden, wird mittels Fishers Z-Transformation untersucht. Strukturgleichungsmodelle (M*plus*) analysieren, inwiefern die Beurteilung der Schülerinnen und Schüler ihres Unterrichts (Vorbereitung auf das Abitur im Unterricht, Autonomie- und Kompetenzunterstützung, Leistungserwartungen, Motivierungsfähigkeit der Lehrperson), die Halbjahresnote 13/1 sowie das Schulklima die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg beeinflussen (Abb. 1). Erfolgsunsicherheit im Abitur und Angst vor Misserfolg (abhängige Variablen) werden latent, die übrigen Variablen manifest modelliert.

Multiple Gruppenvergleiche mit drei Gruppen (2007, 2008, 2011) prüfen, ob die Effekte kurzfristig (2007-2008) und/oder längerfristig (2007-2011) zwischen dezentralen und zentralen Abiturprüfungen differieren. Die Prüfung auf Messinvarianz zwischen den Jahren ergibt für sämtliche Analysen (mindestens partielle) skalare Invarianz. Im Anschluss werden schrittweise zunächst die Regressionskoeffizienten der latenten Variablen auf die manifesten Variablen, dann die Kovarianzen der latenten Variablen, die Korrelationen zwischen den latenten Variablen und schließlich die Korrelationen zwischen den manifesten Variablen über die Jahre restringiert. Ist dies jeweils ohne signifikante Verschlechterung des Modellfits im Vergleich zum weniger restriktiven Modell möglich, unterscheiden sich die Koeffizienten zwischen den Jahren nicht signifikant (vgl. Christ & Schlüter, 2012).

Im Unterschied zu den deskriptiven Analysen werden im Strukturgleichungsmodell lediglich drei der fünf Items der Skala Erfolgsunsicherheit im Abitur verwendet. Den Berechnungen der

Strukturgleichungsmodelle vorangestellte explorative Strukturgleichungsmodelle mit multiplen Gruppenvergleichen zur Ermittlung der Anzahl und Struktur der Faktoren legen den Ausschluss zweier Items aufgrund von Nebenladungen nahe.

Die Berechnung der *Intraclass Correlation* (ICC) zur Bestimmung des Verhältnisses der Varianz auf Level 2 im Verhältnis zur Gesamtvarianz ergibt in jedem Jahr jeweils geringe Werte: Erfolgsunsicherheit im Abitur: 2007: ICC = 0.025, 2008: ICC = 0.043, 2011: ICC = 0.031; Angst vor Misserfolg: 2007: ICC = 0.011, 2008: ICC = 0.006, 2011: ICC = 0.016. Da zudem die Anzahl der Einheiten auf Level 2 mit 17 Schulen deutlich hinter den Faustregeln (vgl. Scherbaum & Ferreter, 2009) zurückbleibt, wird auf eine mehrebenenanalytische Auswertung verzichtet.

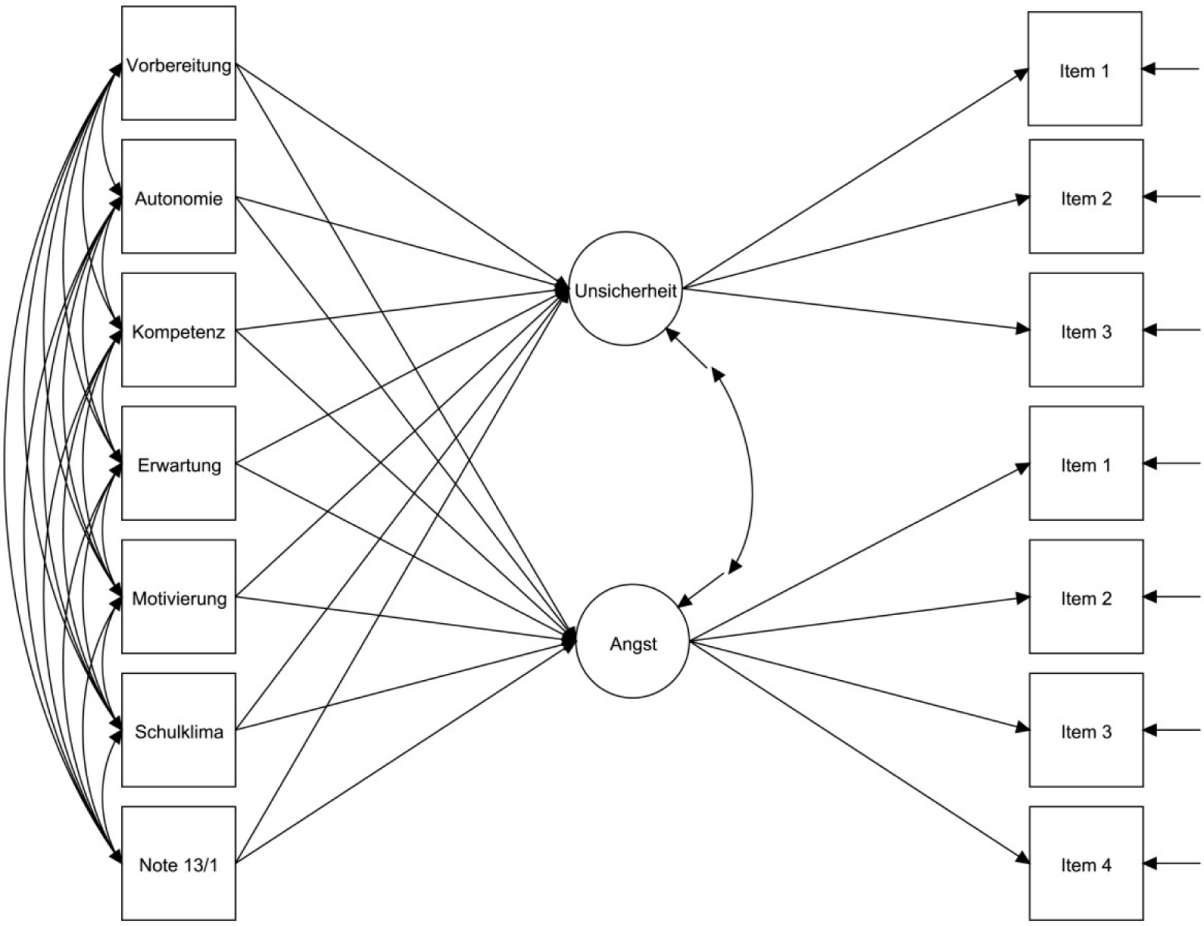


Abbildung 1: Strukturgleichungsmodell der Einflüsse von Vorbereitung im Unterricht, Autonomieunterstützung, Kompetenzunterstützung, Leistungserwartungen, Motivierungsfähigkeit, Schulklima und Halbjahresnote 13/1 auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg

5. BEFUNDE

Deskriptive Auswertungen und Jahresvergleiche (Tab. 2) vermitteln einen ersten Eindruck der Emotionen über die Jahre (Forschungsfrage 1). Die Erfolgsunsicherheit im Abitur liegt in allen Jahren etwas oberhalb des arithmetischen Mittels von 2.5. Im ersten Jahr der zentralen Abiturprüfungen 2008 fühlen sich die Schülerinnen und Schüler unsicherer als unter dezentralen Prüfungen 2007 ($\kappa = .10^{***}$; $d = 0.15$)[1], 2011 ist das Niveau von 2007 wieder erreicht ($\kappa = .01$ n.s.; $d = 0.06$).

Die Angst vor Misserfolg fällt im Vergleich dazu jeweils geringer aus und unterscheidet sich zwischen den Jahren nicht (2007-2008: $\kappa = .04$ n.s.; $d = 0.05$; 2007-2011: $\kappa = -.01$ n.s.; $d = -0.05$).

Tabelle 2: Gepoolte Mittelwerte, gepoolte Standardabweichungen und Standardfehler der Mittelwertschätzung der Variablen sowie Differenzen zwischen den Jahren

	2007			2008			2011			Jahres-
	M	S	SE	M	S	SE	M	S	SE	differenzen
Erfolgs-unsicherheit	2.54	0.70	0.02	2.64	0.67	0.02	2.58	0.70	0.02	07-08: *** 07-11: n.s.
Angst vor Misserfolg	2.12	0.71	0.03	2.15	0.70	0.02	2.08	0.67	0.02	07-08: n.s. 07-11: n.s.
Vorbereitung im Unterricht	2.74	0.75	0.02	2.77	0.73	0.02	2.98	0.77	0.02	07-08: n.s. 07-11: ***
Autonomie-unterstützung	2.45	0.66	0.02	2.47	0.65	0.02	2.52	0.64	0.02	07-08: n.s. 07-11: *
Kompetenz-unterstützung	2.44	0.69	0.02	2.51	0.69	0.02	2.58	0.66	0.02	07-08: ** 07-11: ***
Leistungs-erwartungen	3.01	0.74	0.02	3.09	0.68	0.02	3.03	0.69	0.02	07-08: ** 07-11: n.s.
Motivierungs-fähigkeit	2.52	0.74	0.02	2.60	0.71	0.02	2.62	0.74	0.02	07-08: * 07-11: **
Schulklima	3.52	0.62	0.02	3.51	0.57	0.02	3.51	0.53	0.01	07-08: n.s. 07-11: n.s.
Note 13/1[a]	9.36	2.83	0.08	9.30	2.84	0.07	9.41	2.90	0.07	07-08: n.s. 07-11: n.s.

Anmerkungen: 2007: n = 1368; 2008: n = 1497; 2011: n = 1586; $n_{gesamt} = 4451$; n.s.: nicht signifikant; *: $p < .05$; **: $p < .01$; ***: $p < .001$; [a] Originaldaten: 2007: n = 1350; 2008: n = 1457; 2011: n = 1552; $n_{gesamt} = 4359$

Korrelationen belegen einen über die Zeit konstant hohen Zusammenhang zwischen beiden Emotionen: 2007: $r = .51^{***}$; 2008: $r = .53^{***}$; 2011: $r = .52^{***}$.

[1] Jeweils standardisierte Schätzer und Signifikanzniveau (n.s.: nicht signifikant; +: $p < .10$; *: $p < .05$; **: $p < .01$; ***: $p < .001$).

Weiterhin empfinden die Lernenden im ersten Jahr des Zentralabiturs 2008, dass die Lehrpersonen höhere Leistungen erwarten (κ = .09**; d = 0.12). 2011 sind sie wieder auf dem Niveau von 2007 (κ = .01 n.s.; d = 0.03). 2008 und 2011 werden die Kompetenzunterstützung (2007-2008: κ = .07**; d = 0.10; 2007-2011: κ = .04***; d = 0.21) und die Motivierungsfähigkeit (2007-2008: κ = .07*; d = 0.10; 2007-2011: κ = .02**; d = 0.13) stärker wahrgenommen als 2007. Bezüglich der Vorbereitung im Unterricht (κ = .03 n.s.; d = 0.04) sowie der Autonomieunterstützung (κ = .02 n.s.; d = 0.03) besteht zwischen dezentralen Prüfungen und dem ersten Jahr mit Zentralabitur 2008 kein Unterschied. Mit der Zeit fühlen sich die Lernenden besser vorbereitet (κ = .06***; d = 0.31) und in ihrer Autonomie unterstützt (κ = .02*; d = 0.09). Über alle Jahre hinweg konstant bleiben die Einschätzung des Schulklimas (2007-2008: κ = -.01 n.s.; d = -0.02; 2007-2011: κ = -.00 n.s.; d = -0.02) und die Halbjahresnoten 13/1 (2007-2008: κ = -.05 n.s.; d = -0.02; 2007-2011: κ = .01 n.s.; d = 0.02). Insgesamt fallen die Differenzen gering aus, sodass die Abiturientinnen und Abiturienten die verschiedenen Aspekte ihres persönlichen Befindens, des Unterrichts und des Schulklimas über die Jahre hinweg relativ stabil bewerten.

Zusätzlich werden für die Analyse differenzieller Wirkungen in Abhängigkeit der Vorleistung die deskriptiven Statistiken für jedes Leistungsniveau separat berechnet (Forschungsfrage 3; vgl. Tab. 3-5).

In allen Dimensionen zeigen sich signifikante Unterschiede zwischen den drei Leistungsgruppen. Lernende des unteren Leistungsniveaus weisen die höchsten Werte bei Erfolgsunsicherheit im Abitur, Angst vor Misserfolg und Leistungserwartungen der Lehrpersonen auf bei gleichzeitiger negativerer Bewertung der Unterrichtsgestaltung und des Schulklimas. Bei den leistungsstarken Schülerinnen und Schüler ist das Gegenteil der Fall, während die mittlere Leistungsgruppe eine Zwischenposition einnimmt.

Bei Betrachtung der Jahresvergleiche wird deutlich, dass sich das Ausmaß an Angst vor Misserfolg nicht signifikant verändert. Im ersten Jahr zentraler Abiturprüfungen 2008 empfinden diejenigen des mittleren und oberen Leistungsniveaus signifikant mehr Erfolgsunsicherheit im Abitur als 2007 (mittleres: κ = .11**; d = 0.17; oberes: κ = .14**; d = 0.29). Auf die untere Leistungsgruppe trifft das nicht zu (κ = .05 n.s.; d = 0.09). Im selben Zeitraum nehmen Abiturientinnen und Abiturienten des unteren und oberen Leistungsniveaus signifikant erhöhte Erwartungen der Lehrpersonen wahr, die des mittleren hingegen nicht (unteres: κ = .09*; d = 0.14; mittleres: κ = .05 n.s.; d = 0.06; oberes: κ = .14*; d = 0.19). Alle drei

Gruppen schätzen die Vorbereitung auf das Abitur im Unterricht 2011 signifikant positiver ein als 2007 (unteres: $\kappa = .06^{***}$; d = 0.33; mittleres: $\kappa = .06^{***}$; d = 0.32; oberes: $\kappa = .05^{***}$; d = 0.28), ebenso die Kompetenzunterstützung in 2008 und 2011 (2008: unteres: $\kappa = .11^{*}$; d = 0.16; mittleres: $\kappa = .01$ n.s.; d = 0.01; oberes: $\kappa = .17^{***}$; d = 0.28; 2011: unteres: $\kappa = .03^{*}$; d = 0.18; mittleres: $\kappa = .02^{*}$; d = 0.15; oberes: $\kappa = .06^{***}$; d = 0.40).

Tabelle 3: Gepoolte Mittelwerte, gepoolte Standardabweichungen und Standardfehler der Mittelwertschätzung der Variablen sowie Differenzen zwischen den Jahren, unteres Quartil

	2007			2008			2011			Jahres-
	M	S	SE	M	S	SE	M	S	SE	differenzen
Erfolgs-unsicherheit	2.89	0.62	0.04	2.94	0.59	0.03	2.96	0.61	0.03	07-08: n.s. 07-11: n.s.
Angst vor Misserfolg	2.31	0.72	0.04	2.38	0.69	0.04	2.29	0.66	0.03	07-08: n.s. 07-11: n.s.
Vorbereitung im Unterricht	2.62	0.76	0.04	2.71	0.74	0.04	2.87	0.78	0.04	07-08: n.s. 07-11: ***
Autonomie-unterstützung	2.32	0.65	0.04	2.36	0.61	0.03	2.35	0.63	0.03	07-08: n.s. 07-11: n.s.
Kompetenz-unterstützung	2.13	0.69	0.04	2.24	0.65	0.03	2.26	0.67	0.03	07-08: * 07-11: *
Leistungs-erwartungen	3.19	0.68	0.04	3.27	0.59	0.03	3.18	0.67	0.03	07-08: * 07-11: n.s.
Motivierungs-fähigkeit	2.41	0.69	0.04	2.46	0.66	0.03	2.48	0.70	0.03	07-08: n.s. 07-11: n.s.
Schulklima	3.44	0.63	0.04	3.41	0.62	0.03	3.41	0.54	0.03	07-08: n.s. 07-11: n.s.
Note 13/1[a]	5.84	1.23	0.06	5.83	1.20	0.06	5.85	1.17	0.06	07-08: n.s. 07-11: n.s.

Anmerkungen: 2007: n = 372; 2008: n = 422; 2011: n = 437; n_{gesamt} = 1231; n.s.: nicht signifikant; *: $p < .05$; ***: $p < .001$;
[a] Originaldaten

Zusätzlich finden sich bei den leistungsstarken Lernenden weitere Differenzen zwischen den Jahren beim Autonomieerleben (2007-2011: $\kappa = .03^{*}$; d = 0.19) und der Motivierungsfähigkeit der Lehrpersonen (2007-2008: $\kappa = .20^{***}$; d = 28; 2007-2011: $\kappa = .04^{**}$; d = 0.22), sodass in dieser Gruppe die meisten Veränderungen auftreten.

Insgesamt fallen die Effekte zwischen den Jahren innerhalb der einzelnen Gruppen stärker aus als in der Gesamtstichprobe.

Tabelle 4: Gepoolte Mittelwerte, gepoolte Standardabweichungen und Standardfehler der Mittelwertschätzung der Variablen sowie Differenzen zwischen den Jahren, mittlere Quartile

	2007			2008			2011			Jahres-
	M	S	SE	M	S	SE	M	S	SE	differenzen
Erfolgs-unsicherheit	2.54	0.65	0.03	2.64	0.62	0.02	2.58	0.63	0.02	07-08: ** 07-11: n.s.
Angst vor Misserfolg	2.12	0.69	0.03	2.17	0.68	0.03	2.06	0.66	0.03	07-08: n.s. 07-11: n.s.
Vorbereitung im Unterricht	2.76	0.74	0.03	2.73	0.72	0.03	3.00	0.74	0.03	07-08: n.s. 07-11: ***
Autonomie-unterstützung	2.48	0.65	0.03	2.46	0.65	0.03	2.53	0.64	0.02	07-08: n.s. 07-11: n.s.
Kompetenz-unterstützung	2.48	0.64	0.03	2.49	0.66	0.03	2.57	0.59	0.02	07-08: n.s. 07-11: *
Leistungs-erwartungen	3.02	0.73	0.03	3.07	0.68	0.03	3.00	0.69	0.03	07-08: n.s. 07-11: n.s.
Motivierungs-fähigkeit	2.56	0.74	0.03	2.59	0.73	0.03	2.63	0.74	0.03	07-08: n.s. 07-11: n.s.
Schulklima	3.52	0.62	0.03	3.51	0.55	0.02	3.53	0.52	0.02	07-08: n.s. 07-11: n.s.
Note 13/1[a]	9.50	1.11	0.04	9.53	1.09	0.04	9.43	1.10	0.04	07-08: n.s. 07-11: n.s.

Anmerkungen: 2007: n = 648; 2008: n = 682; 2011: n = 693; n_{gesamt} = 2023; n.s.: nicht signifikant; *: p < .05; **: p < .01; ***: p < .001; [a] Originaldaten

Tabelle 5: Gepoolte Mittelwerte, gepoolte Standardabweichungen und Standardfehler der Mittelwertschätzung der Variablen sowie Differenzen zwischen den Jahren, oberes Quartil

	2007			2008			2011			Jahres-
	M	S	SE	M	S	SE	M	S	SE	differenzen
Erfolgs-unsicherheit	2.12	0.65	0.04	2.27	0.67	0.04	2.18	0.69	0.03	07-08: ** 07-11: n.s.
Angst vor Misserfolg	1.86	0.65	0.04	1.85	0.66	0.04	1.89	0.66	0.03	07-08: n.s. 07-11: n.s.
Vorbereitung im Unterricht	2.85	0.73	0.05	2.95	0.72	0.04	3.06	0.77	0.04	07-08: + 07-11: ***
Autonomie-unterstützung	2.58	0.67	0.04	2.65	0.66	0.04	2.70	0.58	0.03	07-08: n.s. 07-11: *
Kompetenz-unterstützung	2.73	0.61	0.04	2.90	0.61	0.03	2.96	0.54	0.03	07-08: *** 07-11: ***
Leistungs-erwartungen	2.78	0.77	0.05	2.92	0.73	0.04	2.91	0.71	0.03	07-08: * 07-11: n.s.
Motivierungs-fähigkeit	2.60	0.76	0.05	2.80	0.71	0.04	2.76	0.74	0.04	07-08: ** 07-11: **
Schulklima	3.63	0.60	0.04	3.62	0.55	0.03	3.60	0.51	0.02	07-08: n.s. 07-11: n.s.
Note 13/1[a]	13.05	0.99	0.05	13.00	1.01	0.05	13.07	1.00	0.05	07-08: n.s. 07-11: n.s.

Anmerkungen: 2007: n = 330; 2008: n = 353; 2011: n = 422; n_{gesamt} = 1231; n.s.: nicht signifikant; +: p <.10; *: p < .05; **: p < .01; ***: p < .001; [a] Originaldaten

Strukturgleichungsmodell

Das Gesamtmodell (Abb. 1) weist einen sehr guten Modellfit auf: $\chi^2 = 670.698$; df = 234; $\chi^2/df = 3.17$; RMSEA = 0.04 (0.032/0.039); CFI = 0.96; TLI = 0.97; SRMR = 0.04. Die Werte der einzelnen Koeffizienten sind in Tabelle 6 aufgeführt.

Das Ausmaß der Erfolgsunsicherheit im Abitur ändert sich kurzfristig ($\alpha = -.23$*), jedoch nicht längerfristig ($\alpha = -.28$ n.s.). Die Angst vor Misserfolg bleibt hingegen stabil (2008: $\alpha = .07$ n.s.; 2011: $\alpha = .02$ n.s.) (Forschungsfrage 1).

Über die Jahre lassen sich fast alle Koeffizienten restringieren ohne dass sich der Modellfit signifikant verschlechtert (Forschungsfrage 2). Konstant wird die Erfolgsunsicherheit im Abitur vom Gefühl einer guten Vorbereitung im Unterricht und der Kompetenzunterstützung gemildert sowie durch hohe Leistungserwartungen der Lehrperson verstärkt. Autonomieunterstützung und Motivierungsfähigkeit sind durchgängig nicht von Bedeutung. Signifikante Unterschiede bestehen in einer kurzfristigen Verringerung des Effekts der Halbjahresnote 13/1 sowie einer längerfristigen Abnahme der Wirkung des Schulklimas.

Tabelle 6: Standardisierte Effekte von Vorbereitung im Unterricht, Autonomieunterstützung, Kompetenzunterstützung, Leistungserwartungen der Lehrperson, Motivierungsfähigkeit, Schulklima sowie Halbjahresnote 13/1 auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg

	2007		2008		2011		Restriktion
Pfad	λ	**SE**	λ	**SE**	λ	**SE**	**über Jahre**
Erfolgsunsicherheit im Abitur							
Vorbereitung im Unterricht	-.17***	.02	-.18***	.02	-.17***	.02	2007-2008 2007-2011
Autonomie-unterstützung	-.00 n.s.	.02	-.00 n.s.	.02	-.00 n.s.	.02	2007-2008 2007-2011
Kompetenz-unterstützung	-.22***	.02	-.24***	.02	-.23***	.02	2007-2008 2007-2011
Leistungserwartungen	.15***	.02	.15***	.02	.14***	.02	2007-2008 2007-2011
Motivierungsfähigkeit	.02 n.s.	.02	.03 n.s.	.02	.02 n.s.	.02	2007-2008 2007-2011
Schulklima	-.08***	.02	-.08***	.02	.01 n.s.	.02	2007-2008
Note 13/1[a]	-.44***	.02	-.32***	.02	-.45***	.02	2007-2011
Angst vor Misserfolg							
Vorbereitung im Unterricht	-.12***	.02	-.12***	.02	-.12***	.02	2007-2008 2007-2011
Autonomie-unterstützung	.05*	.02	.05*	.02	.05*	.02	2007-2008 2007-2011
Kompetenz-unterstützung	-.18***	.03	-.18***	.03	-.18***	.03	2007-2008 2007-2011
Leistungserwartungen	.19***	.02	.18***	.02	.18***	.02	2007-2008 2007-2011
Motivierungsfähigkeit	.07**	.02	.07**	.03	.07**	.02	2007-2008 2007-2011
Schulklima	-.05*	.02	-.04*	.02	-.04*	.02	2007-2008 2007-2011
Note 13/1[a]	-.18***	.02	-.18***	.02	-.18***	.02	2007-2008 2007-2011

Anmerkungen: 2007: n = 1363; 2008: n = 1483; 2011: n = 1574; n_{gesamt} = 4420; n.s.: nicht signifikant; *: p < .05; **: p < .01; ***: p < .001; [a] Originaldaten

Die Effekte für die Angst vor Misserfolg sind über alle Jahre hinweg stabil. Das Gefühl einer guten Vorbereitung im Unterricht, Kompetenzunterstützung, ein positives Schulklima sowie eine gute Note im Halbjahr 13/1 reduzieren die Angst vor Misserfolg, während hohe Leistungserwartungen der Lehrperson, Autonomieunterstützung und Motivierungsfähigkeit sie verstärken.

Über die Jahre zeigen sich kurz- und längerfristige Differenzen in den Emotionen zwischen den Leistungsgruppen (Forschungsfrage 3). Die Erfolgsunsicherheit im Abitur leistungsschwächerer Lernender unterscheidet sich 2007 und 2008 nicht ($\alpha = 0.07$ n.s.), steigt jedoch tendenziell in 2011 ($\alpha = 0.11^{+}$). Die Angst vor Misserfolg nimmt im ersten Jahr zentraler Abiturprüfungen 2008 zu ($\alpha = 0.17^{*}$), ist 2011 jedoch wieder auf dem Niveau von 2007 ($\alpha = -0.02$ n.s.).

Diejenigen des mittleren Leistungsniveaus empfinden hingegen 2008 signifikant weniger Erfolgsunsicherheit im Abitur als 2007 ($\alpha = -0.42^{*}$). 2007 und 2011 unterscheiden sich nicht ($\alpha = 0.08$ n.s.). Die Angst vor Misserfolg differiert nicht zwischen den Jahren (2008: $\alpha = 0.06$ n.s.; 2011: $\alpha = -0.03$ n.s.).

Im Gegensatz dazu verspüren leistungsstarke Schülerinnen und Schüler 2008 signifikant mehr Erfolgsunsicherheit im Abitur ($\alpha = 0.38^{***}$) und 2011 signifikant weniger ($\alpha = -1.28^{***}$). Die Angst vor Misserfolg steigt tendenziell von 2007 zu 2011 (2008: $\alpha = 0.03$ n.s.; 2011: $\alpha = 0.15^{+}$).

Die für jedes Leistungsniveau separat berechneten Strukturgleichungsmodelle weisen jeweils ebenfalls einen sehr guten Modellfit auf (unteres: $\chi^2 = 446.60$; df = 240; $\chi^2/df = 1.86$; RMSEA = 0.04 (0.034/0.046); CFI = 0.93; TLI = 0.94; SRMR = 0.06; mittleres: $\chi^2 = 475.51$; df = 238; $\chi^2/df = 2.00$; RMSEA = 0.03 (0.028/0.036); CFI = 0.96; TLI = 0.96; SRMR = 0.04; oberes: $\chi^2 = 479.41$; df = 238; $\chi^2/df = 2.01$; RMSEA = 0.04 (0.038/0.049); CFI = 0.93; TLI = 0.94; SRMR = 0.06). Bis auf zwei Ausnahmen lassen sich alle Koeffizienten ohne signifikante Verschlechterung des Modellfits über die Jahre restringieren. Auf die Erfolgsunsicherheit im Abitur wirken sich alle unabhängigen Variablen analog zum Gesamtmodell aus. Einzig beim Schulklima bestehen Unterschiede zwischen den Gruppen: Während es beim unteren Leistungsniveau keine Rolle spielt, reduziert es bei den Schülerinnen und Schülern des mittleren Leistungsniveaus in allen Jahren die Erfolgsunsicherheit im Abitur signifikant. Gleiches gilt für die leistungsstarken Abiturientinnen und Abiturienten für 2007 und 2008, nicht jedoch für 2011

(Umkehrung des Effekts). Mehr Differenzen zwischen den Leistungsniveaus finden sich bei der Angst vor Misserfolg, wobei die Wirkungsrichtungen der Kompetenzunterstützung und der Leistungserwartungen der Lehrpersonen für alle drei Leistungsniveaus und der Vorbereitung im Unterricht, Motivierungsfähigkeit, Halbjahresnote 13/1 und des Schulklimas für jeweils zwei Leistungsgruppen identisch zum Gesamtmodell ausfallen. Die Steigerung der Angst vor Misserfolg durch Autonomieunterstützung im Gesamtmodell geht hingegen einzig auf das mittlere Leistungsniveau zurück. Die meisten Abweichungen zum Gesamtmodell weist die obere Leistungsgruppe auf.

Insgesamt bestehen zwischen der leistungsstarken, mittleren und leistungsschwächeren Gruppe sowohl Unterschiede in der Ausprägung des emotionalen, unterrichtlichen und schulischen Erlebens als auch hinsichtlich der kurz- und längerfristigen Entwicklung der Emotionen sowie bezüglich der Wirkungen unterrichtlicher und schulischer Merkmale auf die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg.

6. DISKUSSION

Legt man zur Beantwortung der ersten Forschungsfrage, wie sich die *Emotionen in kurz- und längerfristiger Perspektive* gestalten und ob Veränderungen in Zusammenhang mit der Implementation des Zentralabiturs auftreten, die Ergebnisse der latenten Modellierung zugrunde, fällt die Erfolgsunsicherheit im Abitur im ersten Jahr mit Zentralabitur 2008 und erreicht 2011 wieder das Niveau von 2007. Die Angst vor Misserfolg bleibt dagegen über die Jahre konstant. Somit ist die Hypothese einer kurzfristigen Verstärkung und einer längerfristigen Reduktion für beide Emotionen abzulehnen (analog 2007-2009: vgl. Maag Merki, 2012). Die von Jürges et al. (2009) für das Ende der Sekundarstufe I berichteten stärkeren negativen Emotionen unter den Bedingungen zentraler Abschlussprüfungen zeigen sich am Ende der Sekundarstufe II nicht. Es ist denkbar, dass 2007 aufgrund der zentralen Abiturprüfungen in den Grundkursen auch in den Leistungskursen trotz dezentraler Prüfungen bereits ein größeres Maß an Besorgtheit und Unsicherheit vorhanden war und daher die Differenzen zwischen den Jahren gering ausfallen. Diesen Transfereffekt müssten weitere Analysen prüfen.

Bezüglich der *Wirkung schulischer und unterrichtlicher Faktoren auf die Emotionen* (Forschungsfrage 2) zeigt sich unabhängig von der Prüfungsform in allen Jahren: Je höher die Vorbereitung im Unterricht auf die Abiturprüfungen, die Kompetenzunterstützung und die Halbjahresnote 13/1 ausfallen, desto geringer sind die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg. Verstärkt werden beide Emotionen durch hohe Leistungserwartungen der Lehrperson, die Angst vor Misserfolg zusätzlich durch Autonomieunterstützung und Motivierungsfähigkeit. Der Zusammenhang von der Wahrnehmung des sozialen und didaktischen Kontextes der Aufgabe, dafür erforderlicher Kompetenzen und Emotionen schließt an das *Adaptable Learning Model* (Boekaerts, 1992) an und bestätigt die theoretisch postulierte Interaktion von Person und (schulischer) Umwelt (vgl. Deci & Ryan, 1993; Zeidner, 1998). Zudem decken sich die Befunde mit denen anderer Studien dahingehend, dass hohe Leistungserwartungen Angst vor Prüfungen hervorrufen sowie dass eine gute Vorbereitung, Kompetenzerleben und ein positives Schulklima negative Emotionen mildern (vgl. Frenzel et al., 2007; Gläser-Zikuda & Fuß, 2008; Hospel & Galand, 2016; Ledergerber, 2015; Oerke, 2012; Reyes et al., 2012) und erweitern deren Geltung für Lernende am Ende der Sekundarstufe II.

Die Hypothese, dass die Implementation zentraler Abiturprüfungen den Einfluss schulischer und unterrichtlicher Faktoren auf die Emotionen kurz- und längerfristig verstärkt, bestätigt sich bis auf zwei Ausnahmen nicht. Die mit dem Zentralabitur einhergehende gesteigerte Bedeutung des den Prüfungen vorgelagerten Unterrichts (Oerke et al., 2011) lässt sich im vorliegenden Fall nicht durch vergrößerte Zusammenhänge abbilden. Dies könnte, in Kombination mit dem Befund, dass beide Emotionen über die Jahre stabil bleiben, darin begründet sein, dass Abschlussprüfungen, ob dezentral oder zentral, für Schülerinnen und Schüler stets einen *high stakes*-Charakter besitzen. Deren Organisation scheint lediglich von untergeordneter Bedeutung zu sein. Dass die Abiturientinnen und Abiturienten über die Zeit im Sinne der Unterrichtsqualität positive Veränderungen in der Unterrichtsgestaltung wahrnehmen, zeigen die deskriptiven Statistiken. Einerseits könnte der mit den zentralen Prüfungen einhergehende Druck und die Unsicherheiten für die Lehrpersonen (vgl. Bishop, 1999) dazu führen, dass diese mehr in die Unterrichtsgestaltung investieren (vgl. Jones & Egley, 2004; Jürges et al., 2009). Andererseits ist mit den gewonnenen Erfahrungen und den verfügbaren Materialien (z. B. Abiturprüfungen vergangener Jahre) die Vorbereitung auf das Abitur, auch im Unterricht, gezielter möglich, sodass der Fokus auf andere Aspekte des Unterrichts gelegt werden kann (vgl. Klein, 2016; Maag Merki & Oerke, 2017; allgemeiner:

Thiel, 2016). Dies wiederum könnte sich auf den Selbstwert, die Motivation, das Engagement oder die Erwartung von (Miss-)Erfolg auswirken und so zur Verringerung von Unsicherheit und Besorgnis beitragen (vgl. Deci & Ryan, 1993; Hospel & Galand, 2016; Pekrun, 2006; Peters-Häderle, 2006).

Die *Differenzierung nach Leistungsniveau* (Forschungsfrage 3) offenbart hypothesenkonform in allen Jahren bei Lernenden des unteren Quartils eine ungünstigere Einschätzung der unterrichtlichen und schulischen Dimensionen bei höherer Erfolgsunsicherheit im Abitur und Angst vor Misserfolg als bei denjenigen der beiden anderen Leistungsgruppen (vgl. Gläser-Zikuda & Mayring, 2003; Goetz et al., 2004). Das heißt jedoch nicht, dass Lernende des mittleren und oberen Leistungsniveaus diese Emotionen nicht empfinden (vgl. Gläser-Zikuda & Mayring, 2003; Ryan et al., 2007; Zeidner & Schleyer, 1998). Die kurz- und längerfristige Entwicklung der Erfolgsunsicherheit im Abitur unterscheidet sich zwar zwischen den Gruppen, jedoch nicht hypothesenkonform. Bei latenter Modellierung variiert die Erfolgsunsicherheit im Abitur von 2007 zu 2008 zwischen keine Veränderung (unteres Quartil), Abnahme (mittlere Quartile) und Zunahme (oberes Quartil). Längerfristig steigt sie bei den leistungsschwächeren Abiturientinnen und Abiturienten, bleibt beim mittleren Quartil unverändert und reduziert sich bei den Leistungsstarken. Die Angst vor Misserfolg bleibt in den mittleren Quartilen stabil, wohingegen sie längerfristig beim oberen Quartil tendenziell zunimmt. Die leistungsschwächeren Schülerinnen und Schüler empfinden 2008 eine höhere Angst vor Misserfolg als 2007, nicht jedoch 2011. Die Hypothese ungünstigerer Entwicklungen der Emotionen des unteren Quartils bestätigt sich demnach lediglich für die Erfolgsunsicherheit im Abitur.

Zwar liegt der Fokus beider Emotionen auf den Ergebnissen (*outcome emotions*; Pekrun, 2006), dennoch ist er verschieden: Die Items zur Erfolgsunsicherheit im Abitur beziehen sich konkret auf die bevorstehenden Prüfungen, sind also eher situationsspezifisch (*state*), wohingegen die Items zur Angst vor Misserfolg allgemeiner formuliert sind. Diese kann bereits seit längerem bestehen (*trait*) und in den bisherigen (Lern-)Erfahrungen begründet sein (vgl. Zeidner, 1998). Gläser-Zikuda und Mayring (2003, S. 117-118) zufolge reflektieren leistungsstarke Schülerinnen und Schüler mehr über ihr Lernen und ihre Leistungen, sodass sie häufiger *state*-Angst empfinden, wohingegen leistungsschwächere häufiger *trait*-Angst berichten. Hier ist die Befundlage umgekehrt, wobei weitere Analysen

klären müssten, ob die variierenden Ergebnisse auf das unterschiedliche Alter der Befragten oder verschiedene methodische Zugänge zurückzuführen sind.

Entsprechend der Annahme differenzieller Wirkungen unterrichtlicher und schulischer Merkmale auf die Emotionen in Abhängigkeit der Halbjahresnote 13/1 unterscheiden sich die drei Leistungsgruppen in einigen Dimensionen, in anderen nicht. Veränderungen über die Jahre zeigen sich (fast) nicht. Demnach wirkt sich unabhängig von der Organisation der Abiturprüfungen die Lernumgebung je nach Leistungsstand zumindest teilweise unterschiedlich aus (vgl. Muijs et al., 2005; Vanlaar et al., 2016). Eine Gemeinsamkeit ist, dass hohe Leistungserwartungen der Lehrpersonen die Emotionen verstärken (vgl. Frenzel et al., 2007; Goetz et al., 2004; Schumacher, 2016); gleichzeitig werden sie, zumindest größtenteils, durch gute Vorbereitung im Unterricht und Kompetenzunterstützung gemildert (vgl. Deci & Ryan, 1993; Hospel & Galand, 2016; Ledergerber, 2015). Das Autonomieerleben spielt entgegen der Selbstbestimmungstheorie von Deci und Ryan (1993) und Befunden, dass das Autonomieerleben negativ mit negativen Emotionen verbunden ist (vgl. Hospel & Galand, 2016), mit einer Ausnahme (mittleres Leistungsniveau: Angst vor Misserfolg) keine Rolle.

Die Implementation des Zentralabiturs auf der Makroebene durchzieht mittels Rekontextualisierungsprozessen (vgl. Fend, 2008) die Meso- und die Mikroebene, sodass es die Lehrpersonen und ihre Unterrichtsgestaltung beeinflusst (vgl. Jones & Egley, 2004; Klein, 2016; Maag Merki & Oerke, 2017). Anzunehmen ist, dass sich zudem die mit der Reform des Abiturs einhergehenden Emotionen der Lehrpersonen (vgl. Maué, Maag Merki & Oerke, 2012; Oerke, 2012) direkt und indirekt über den Unterricht auf die Emotionen der Lernenden auswirken (vgl. Jennings & Greenberg, 2009). Somit sind die Emotionen der Schülerinnen und Schüler in ein komplexes Zusammenspiel individueller, unterrichtlicher und schulischer Merkmale eingebettet. Daraus ergeben sich einerseits vielfältige Ansatzpunkte zur Reduktion negativer und zum Aufbau positiver Emotionen unter Berücksichtigung der spezifischen Merkmale jeder Leistungsgruppe (vgl. Jennings & Greenberg, 2009; Peters-Häderle, 2006; Reyes et al., 2012). Andererseits zeigt sich, dass über verschiedene Kohorten hinweg und unter unterschiedlichen Rahmenbedingungen die Emotionen eine hohe Stabilität aufweisen.

7. EINSCHRÄNKUNGEN UND AUSBLICK

Der Beitrag analysiert die Erfolgsunsicherheit im Abitur und die Angst vor Misserfolg über einen Zeitraum von fünf Jahren unter den Rahmenbedingungen der Implementation des Zentralabiturs. Er untersucht den Einfluss und die potenzielle Veränderung schulischer und unterrichtlicher Faktoren auf diese Emotionen, auch in Abhängigkeit des Leistungsniveaus. Es zeigen sich zwar differenzielle Auswirkungen, doch bleibt der Großteil der Koeffizienten über die Jahre stabil. Dies mag im begrenzten Zeitraum (lediglich Daten eines Jahres dezentralen Abiturs) und Untersuchungsausschnitt des komplexen Gefüges von Emotionen, anderen Dimensionen und des Kontextes begründet sein. Der Fokus müsste durch Einbezug persönlicher Merkmale, Motivationen, Attributionen und Selbstkonzepten sowie Emotionen anderer Akteure erweitert werden. Ebenso könnten (fachspezifische) Analysen die Wirkung der Emotionen auf das Abschneiden im Abitur und somit die „Kette" Vorleistung – Emotionen – Abiturleistung in den Blick nehmen (Schnabel, 1998).

Die Befunde gelten vorrangig für Leistungskurse mit einem Wechsel von dezentralen zu zentralen Abiturprüfungen in Bremen. Ob sie sich auf andere (Bundes-)Länder und Kurse übertragen lassen, muss offen bleiben. Dennoch zeigen sie erstens, dass die Befürchtung, die Implementation des Zentralabiturs würde generell mit einem erhöhten Druck und Stress einhergehen, zumindest im vorliegenden Fall unbegründet zu sein scheint. Stattdessen muss zwischen verschiedenen Gruppen von Lernenden differenziert werden. Zweitens wird deutlich, dass, unabhängig vom Leistungsniveau, das Gefühl eines kompetenzunterstützenden Unterrichts, der gut auf das Abitur vorbereitet, negative Emotionen mindern kann. Diese Ressource ist bei der Planung und Umsetzung des ‚alltäglichen Geschäfts' von Schule ebenso wie bei künftigen Reformen zu berücksichtigen.

LITERATUR

Ahmed, W., van der Werf, G., Minnaert, A. & Kuyper, H. (2010). Students' daily emotions in the classroom: intra-individual variability and appraisal correlates. *British Journal of Educational Psychology, 80*(4), 583-597.

Baumert, J., Gruehn, S., Heyn, S., Köller, O. & Schnabel, K. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU). Skalendokumentation.* Berlin: Max-Planck-Institut für Bildungsforschung.

Bishop, J. H. (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review, 6*, 349-398.

Bishop, J. H. & Wößmann, L. (2004). Institutional Effects in a Simple Model of Educational Production. *Education Economics, 12*(1), 17-38.

Boekaerts, M. (1992). The Adaptable Learning Process: Initiating and Maintaining Behavioural Change. *Applied Psychology: An International Review, 41*(4), 377-397.

Christ, O. & Schlüter, E. (2012). S*trukturgleichungsmodelle mit Mplus. Eine praktische Einführung.* München: Oldenburg.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NY: Erlbaum.

Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik, 39*(2), 223-238.

Eder, F. (1998). *Linzer Fragebogen zum Schul- und Klassenklima für die 8. - 13. Klasse (LFSK 8-13).* Göttingen: Hogrefe.

Fend, H. (2008). S*chule gestalten: Systemsteuerung, Schulentwicklung und Unterrichtsqualität.* Wiesbaden: Verlag für Sozialwissenschaften

Fend, H. (2011). Die Wirksamkeit der Neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. Z*eitschrift für Bildungsforschung, 1*(1), 5-24.

Frenzel, A. C., Goetz, T. & Pekrun, R. (2015). Emotionen. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie (*2., vollst. überarb. & akt. Aufl., S. 201-224). Berlin / Heidelberg: Springer.

Frenzel, A. C., Pekrun, R. & Goetz, T. (2007). Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction, 17*(5), 478-493.

Gläser-Zikuda, M. & Fuß, S. (2008). Lehrerkompetenzen und Schüleremotionen: Wie nehmen Lernende ihre Lehrkräfte emotional wahr? In M. Gläser-Zikuda & J. Seifried (Hrsg.), *Lehrerexpertise. Analyse und Bedeutung unterrichtlichen Handelns* (S. 113-142). Münster: Waxmann.

Gläser-Zikuda, M. & Mayring, P. (2003). A qualitative oriented approach to learning emotions at school. In P. Mayring & C. von Rhöneck (Hrsg.), *Learning Emotions. The Influence of Affective Factors on Classroom Learning* (S. 103-126). Frankfurt a. M. et al.: Peter Lang.

Goetz, T., Pekrun, R., Zirngibl, A., Jullien, S., Kleine, M., Vom Hofe, R. & Blum, W. (2004). Leistung und emotionales Erleben im Fach Mathematik – Längsschnittliche Mehrebenenanalysen. *Zeitschrift für pädagogische Psychologie, 18*(3-4), 201-212.

Grob, U. & Maag Merki, K. (2001). *Überfachliche Kompetenzen. Theoretische Grundlegung und empirische Erprobung eines Indikatorensystems.* Bern: Peter Lang.

Hagenauer, G. & Hascher, T. (2014). Early Adolescents' Enjoyment Experienced in Learning Situations at School and Its Relation to Student Achievement. *Journal of Education and Training Studies, 2*(2), 20-30. doi: 10.11114/jets.v2i2.254

Hargreaves, A. (1998). The emotional practice of teaching. *Teaching and Teacher Education, 14*(8), 835-854.

Heckhausen, J. (1963). *Hoffnung und Furcht in der Leistungsmotivation.* Meisenheim am Glan: Anton Hain.

Hembree, R. (1990). The Nature, Effects, and Relief of Mathematics Anxiety. *Journal for Research in Mathematics Education, 21*(1), 33-46.

Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponenten. *Zeitschrift für pädagogische Psychologie, 5*(2), 121-130.

Hospel, V. & Galand, B. (2016). Are both classroom autonomy support and structure equally important for students' engagement? A multilevel analysis. *Learning and Instruction, 41*, 1-10.

Hugener, I. (2008). *Inszenierungsmuster im Unterricht und Lernqualität. Sichtstrukturen schweizerischen und deutschen Mathematikunterrichts in ihrer Beziehung zu Schülerwahrnehmung und Lernleistung – eine Videoanalyse.* Münster: Waxmann.

Jennings, P. A. & Greenberg, M. T. (2009). The Prosocial Classroom: Teacher Social and Emotional Competence in Relation to Student and Classroom Outcomes. *Review of Educational Research, 79*(1), 491-525.

Jones, B. D. & Egley, R. J. (2004). Voices from the Frontlines: Teachers' Perceptions of High-Stakes Testing. *Education Policy Analysis Archives, 12*(39), 1-29.

Jürges, H., Schneider, K., Senkbeil, M. & Carstensen, C. H. (2009). *Assessment Drives Learning. The Effect of Central Exit Exams on Curricular Knowledge and Mathematical Literacy*. (CESifo Working Paper No. 2666). München: CESifo.

Klein, E. D. (2016). How do teachers prepare their students for statewide exit exams? A comparison of Finland, Ireland, and the Netherlands. *Journal for Educational Research Online, 8*(2), 31-59.

Klein, E. D., Kühn, S. M., van Ackeren, I. & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. *Zeitschrift für Pädagogik, 55*(4), 596-621.

Koretz, D. (2008). Test-based Educational Accountability. Research Evidence and Implications. *Zeitschrift für Pädagogik, 54*(6), 777-790.

Kühn, S. M. (2010). *Steuerung und Innovation durch Abschlussprüfungen?* Wiesbaden: Verlag für Sozialwissenschaften

Ledergerber, C. (2015). *Unterrichtskommunikation und motivational-emotionale Aspekte des Lernens. Eine videobasierte Analyse im Mathematikunterricht*. Münster / New York: Waxmann.

Leutwyler, B. & Maag Merki, K. (2005). *Mittelschulerhebung 2004. Indikatoren zu Kontextmerkmalen gymnasialer Bildung. Perspektive der Schülerinnen und Schüler: Schul- und Unterrichtserfahrungen. Skalen- und Itemdokumentation.* Zürich: Forschungsbereich Schulqualität & Schulentwicklung, Pädagogisches Institut, Universität Zürich.

Lipowsky, F. (2005). Auf den Lehrer kommt es an. Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. In C. Allemann- Ghionda & E. Terhart (Hrsg.), Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern (*51. Beiheft der Zeitschrift für Pädagogik*, S. 47 – 70). Weinheim: Beltz.

Ma, X. (1999). A Meta-Analysis of the Relationship Between Anxiety Toward Mathematics and Achievement in Mathematics. *Journal for Research in Mathematics Education, 30*(5), 520-540.

Maag Merki, K. (2002). *Evaluation Mittelschulen – Überfachliche Kompetenzen. Schlussbericht der ersten Erhebung.* Zürich: Forschungsbereich Schulqualität & Schulentwicklung, Pädagogisches Institut, Universität Zürich.

Maag Merki, K. (2006). *Lernort Gymnasium.* Bern: Haupt.

Maag Merki, K. (2012). Zentrale Prüfungen – empirische Evidenzen der Effekte der Einführung zentraler Abiturprüfungen auf Motivation und Emotion der Schüler/innen. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (S. 247-275). Wiesbaden: VS | Springer.

Maag Merki, K. & Oerke, B. (2012). Methodische Grundlagen der Studie. In K. Maag Merki (Hrsg.), Zentralabitur. *Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 45-61). Wiesbaden: Springer.

Maag Merki, K. & Oerke, B. (2017). Long-term effects of the implementation of state-wide exit exams: a multilevel regression analysis of mediation effects of teaching practices on students' motivational orientations. *Educational Assessment, Evaluation and Accountability, 29*(1), 23-54.

Maier, U. (2002). Eine qualitative Interviewstudie zum Einfluss des Lehrerverhaltens auf Lernemotionen von Schülern im naturwissenschaftlichen Unterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 8*, 85-102.

Maué, E., Maag Merki, K. & Oerke, B. (2012). Emotionales Erleben des Zentralabiturs von Lehrpersonen in Bremen. Längerfristige Effekte der Implementation zentraler Abiturprüfungen. In S. Hornberg & M. Parreira do Amaral (Hrsg.), *Deregulierung im Bildungswesen* (S. 109-130). Münster: Waxmann.

Muijs, D., Campbell, J. & Kyriakides, L. (2005). Making the case for differentiated teacher effectiveness. An overview of research in four key areas. *School effectiveness and school improvement, 16*(1), 51-70.

Muthén, L. K. & Muthén, B. O. (1998-2012). *Mplus User's Guide* (7. Aufl.). Los Angeles, CA: Muthén & Muthén.

Oerke, B. (2012). Emotionaler Umgang von Lehrkräften und Schüler/-innen mit dem Zentralabitur: Unsicherheit, Leistungsdruck und Leistungsattributionen. In K. Maag Merki (Hrsg.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (S. 115-149). Wiesbaden: Springer.

Oerke, B., Maag Merki, K., Holmeier, M. & Jäger, D. J. (2011). Changes in student attributions due to the implementation of central exit exams. E*ducational Assessment, Evaluation and Accountability, 23*(3), 223-241.

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review, 18*(4), 315-341.

Pekrun, R. & Linnenbrink-Garcia, L. (2014). *International Handbook of Emotions in Education*. New York / London: Routledge.

Peters-Häderle, K.-E. (2006). *Erfolgsfurcht und Leistungsangst bei Schülern – eine Trainingsstudie.* Dissertation, Universität Regensburg.

Rakoczy, K., Buff, A. & Lipowsky, F. (2005). Teil 1: Befragungsinstrumente. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis".* Frankfurt a. M.: GEPF & DIPF.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M. & Salovey, P. (2012). Classroom Emotional Climate, Student Engagement, and Academic Achievement. *Journal of Educational Psychology, 104*(3), 700-712.

Roos, A.-L., Bieg, M., Goetz, T., Frenzel, A. C., Taxer, J. & Zeidner, M. (2015). Experiencing more mathematics anxiety than expected? Contrasting trait and state anxiety in high achieving students. *High Ability Studies, 26*(2), 245-258.

Rost, D. H. & Schermer, F. J. (1987). Emotion and Cognition in Coping with Test Anxiety. *Communication & Cognition, 20*(2/3), 225-244.

Rubin, D. B. (1987). *Mutliple Imputation for nonresponse in surveys.* New York: Wiley.

Ryan, K. E., Ryan, A. M., Arbuthnot, K. & Samuels, M. (2007). Students' Motivation for Standardized Math Exams. *Educational Researcher, 36*(1), 5-13.

Scherbaum, C. A. & Ferreter, J. M. (2009). Estimating Statistical Power and Required Sample Sizes for Organizational Research Using Multilevel Modeling. *Organizational Research Methods, 12*(2), 347-367.

Schnabel, K. (1998). *Prüfungsangst und Lernen.* Münster: Waxmann.

Schumacher, C. (2016). *Prüfungsangst in der Schule. Ursachen, Bewältigung und Folgen am Beispiel zentraler Abschlussprüfung.* Münster / New York: Waxmann.

Seifried, J. (2009). *Unterricht aus der Sicht von Handelslehrern.* Frankfurt a. M.: Peter Lang.

Seipp, B. (1990). *Angst und Leistung in Schule und Hochschule. Eine Meta-Analyse.* Frankfurt a. M. et al.: Lang.

Thiel, F. (2016). *Interaktion im Unterricht. Ordnungsmechanismen und Stördynamiken.* Opladen / Toronto: Barbara Budrich.

Urhahne, D. (2015). Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion. *Teaching and Teacher Education, 4*5, 73-82. doi: 10.1016/j.tate.2014.09.006

van Ackeren, I., Klemm, K. & Kühn, S. M. (2015). *Entstehung, Struktur und Steuerung des deutschen Schulsystems. Eine Einführung* (3., überarb. & akt. Aufl.). Wiesbaden: Springer VS.

Vanlaar, G., Kyriakides, L., Panayiotou, A., Vandecandelaere, M., McMahon, L., De Fraine, B. & Van Damme, J. (2016). Do the teacher and school factors of the dynamic model affect high- and low-achieving student groups to the same extent? A cross-country study. *Research Papers in Education, 31*(2), 183-211.

Weiner, B. (2000). Intrapersonal and Interpersonal Theories of Motivation from an Attributional Perspective. *Educational Psychology Review, 12*(1), 1-14.

Wößmann, L. (2003). Zentrale Prüfungen als „Währung" des Bildungssystems: Zur Komplementarität von Schulautonomie und Zentralprüfungen. *Vierteljahrshefte zur Wirtschaftsforschung, 72*(2), 220-237.

Zeidner, M. (1998). *Test anxiety: the state of the art.* New York: Plenum Press.

Zeidner, M. & Schleyer, E. J. (1998). The Big-Fish-Little-Pond Effect for Academic Self-Concept, Test Anxiety, and School Grades in Gifted Children. *Contemporary educational psychology, 24*(4), 305-329.

Zierer, K. (2015). Auf die Lehrperson kommt es an!? Kritisch-konstruktive Betrachtung eines pädagogischen Mythos. In S. Lin-Klitzing, D. Di Fuccia & R. Stengl-Jörns (Hrsg.), *Auf die Lehrperson kommt es an? Beiträge zur Lehrerbildung nach John Hatties „Visible Learning"* (S. 117-126). Bad Heilbrunn: Julius Klinkhardt.

Zimmer, J. W. & Hocevar, D. J. (1994). Effects of massed versus distributed practice of test taking on achievement and test anxiety. *Psychological Reports, 74*(3), 915-919.

LEBENSLAUF

Elisabeth Maué

geboren am 25. Oktober 1984 in Nürnberg, Deutschland

Akademischer Werdegang

2011 – 2018 Doktoratsprogramm Erziehungswissenschaft an der Universität Zürich

2008 – 2010 Master-Studium der Erziehungswissenschaften an der Humboldt-Universität zu Berlin (Schwerpunkt: Internationale Bildungsforschung und Bildungsexpertise)

2005 – 2009 Bachelor-Studium der Erziehungswissenschaften und der Skandinavistik an der Humboldt-Universität zu Berlin (Schwerpunkt: Internationale Bildungsforschung und Bildungsexpertise)

Beruflicher Werdegang

seit 2016 Akademische Mitarbeiterin an der Universität Konstanz,
Fachbereich Wirtschaftswissenschaften, Arbeitsbereich Wirtschaftspädagogik II
(Herr Prof. Dr. Stephan Schumann)
Fachbereich Geschichte und Soziologie, Arbeitsbereich Mikrosoziologie
(Frau Prof. Dr. Claudia Diehl)

2011 – 2016 Wissenschaftliche Assistentin an der Universität Zürich, Institut für Erziehungswissenschaft, Lehrstuhl Theorie und Empirie schulischer Bildungsprozesse
(Frau Prof. Dr. Katharina Maag Merki)

2010 – 2011 Wissenschaftliche Mitarbeiterin an der Humboldt-Universität zu Berlin, Institut für Erziehungswissenschaften, Abteilung Empirische Bildungsforschung und Methodenlehre
(Herr Prof. Dr. Dr. Dr. h.c. Rainer Lehmann)

2008 – 2010 Studentische Hilfskraft an der Humboldt-Universität zu Berlin, Institut für Erziehungswissenschaften, Abteilung Empirische Bildungsforschung und Methodenlehre
(Herr Prof. Dr. Dr. Dr. h.c. Rainer Lehmann)

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Philosophische Fakultät
Studiendekanat
Rämistrasse 69
CH-8001 Zürich
www.phil.uzh.ch

Erklärung

Hiermit erkläre ich, dass die Dissertation von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und diese Dissertation noch an keiner anderen Fakultät eingereicht wurde.

Ort und Datum	Unterschrift
Zürich, 23.01.2018	Elisabeth Mauz